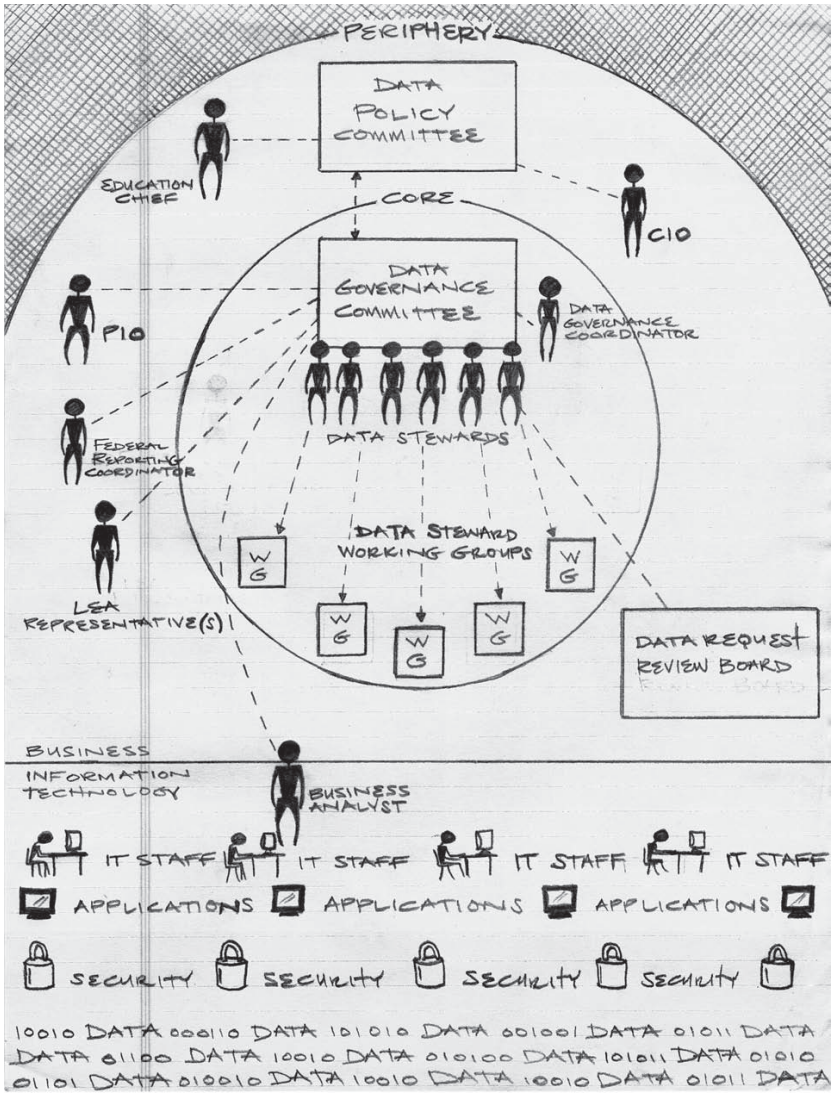
P-20 Data Sharing Policy Workgroup

WORKBOOK



Education Research & Data Center
OFFICE OF FINANCIAL MANAGEMENT
FORECASTING DIVISION

STATE OF WASHINGTON

P-20 Data Sharing Policy Workgroup
WORKBOOK
Table of Contents

5. Meeting Documents
    - Meeting #1 – 3/11/2011
        - Agenda
        - Meeting Notes
        - Handouts
    - Meeting #2 -
        - Agenda
        - Meeting Notes
        - Handouts
    - Meeting #3 -
        - Agenda
        - Meeting Notes
        - Handouts
    - Meeting #4 -
        - Agenda
        - Meeting Notes
        - Handouts
    - Meeting #5 -
        - Agenda
        - Meeting Notes
        - Handouts
    - Meeting #6 -
        - Agenda
        - Meeting Notes
        - Handouts

P-20 Data Sharing Policy Workgroup

MEMBERS (alphabetical listing)

Marc Baldwin
Assistant Director
Office of Financial Management, Forecasting
Executive Sponsor of P-20 Program
Marc.Baldwin@OFM.WA.GOV
360-902-0590

Alice Burman
Assessment Coordinator
Olympia School District
K-12 School District Representative
aburman@osd.wednet.edu
360-596-8542

Bob Butts
Assistant Superintendent, Public Policy & Planning
Office of Superintendent of Public Instruction
Executive Sponsor of P-20 Program
bob.butts@k12.wa.us
360-725-0420

Laura Coghlan
Director, Institutional Research & Assessment
The Evergreen State College
Four-year Institutions Representative
coghlanl@evergreen.edu
360-867-6676

Cynthia Forland
Director, Research & Policy
Employment Security Department
P-20 Partner Representative
CForland@ESD.WA.GOV
360-902-9701

Jan Ignash
Deputy Director, Policy, Planning & Research
Higher Education Coordinating Board
P-20 Partner Representative
jani@hecb.wa.gov
360-704-4168

Tom Jensen
Administrator
Legislative Evaluation & Accountability Program
Executive Sponsor of P-20 Program
jensen.tom@leg.wa.gov
360-786-6111

Joyce Kilmer
ECEAP Program Administrator
Department of Early Learning
ECEAP Provider Representative
joyce.kilmer@del.wa.gov
360-725-2843

Corina McCleary
Chief Information Officer
Department of Early Learning
P-20 Partner Representative
corina.mccleary@del.wa.gov
360-725-4398

Robin Munson
Director, Student Information
Office of Superintendent of Public Instruction
P-20 Partner Representative
Robin.Munson@k12.wa.us
360-725-6356

Mike Reilly
Executive Director
Council of Presidents (postsecondary institutions)
P-20 Partner Representative
mreilly@cop.wsu.edu
360-292-4100

Jim Schmidt
Director, Education Research and Data Center (ERDC)
Office of Financial Management
ERDC Representative
Jim.Schmidt@ofm.wa.gov
360-902-0595

David Stolier
Assistant Attorney General, Education
Washington State Attorney General
Legal Advisor
daves@atg.wa.gov
360-586-0279

Jan Yoshiwara
Director, Education Division
State Board for Community and Technical Colleges
P-20 Partner Representative
jyoshiwara@sbctc.edu
360-704-4353

To be determined
CTC (Two-year) Institutions Representative

WORKGROUP STAFF

Melissa Beard
Senior Forecasting Analyst
Office of Financial Management, ERDC
Melissa.Beard@ofm.wa.gov
360-902-0584

PROJECT MANAGEMENT AND OVERSIGHT

Glenn Briskin
QA Consultant
Briskin Consulting
glenn@briskinconsulting.com
360-561-0897

Deb Came
Data Sharing Project Manager
Office of Financial Management, ERDC
deb.came@ofm.wa.gov
360-902-0491

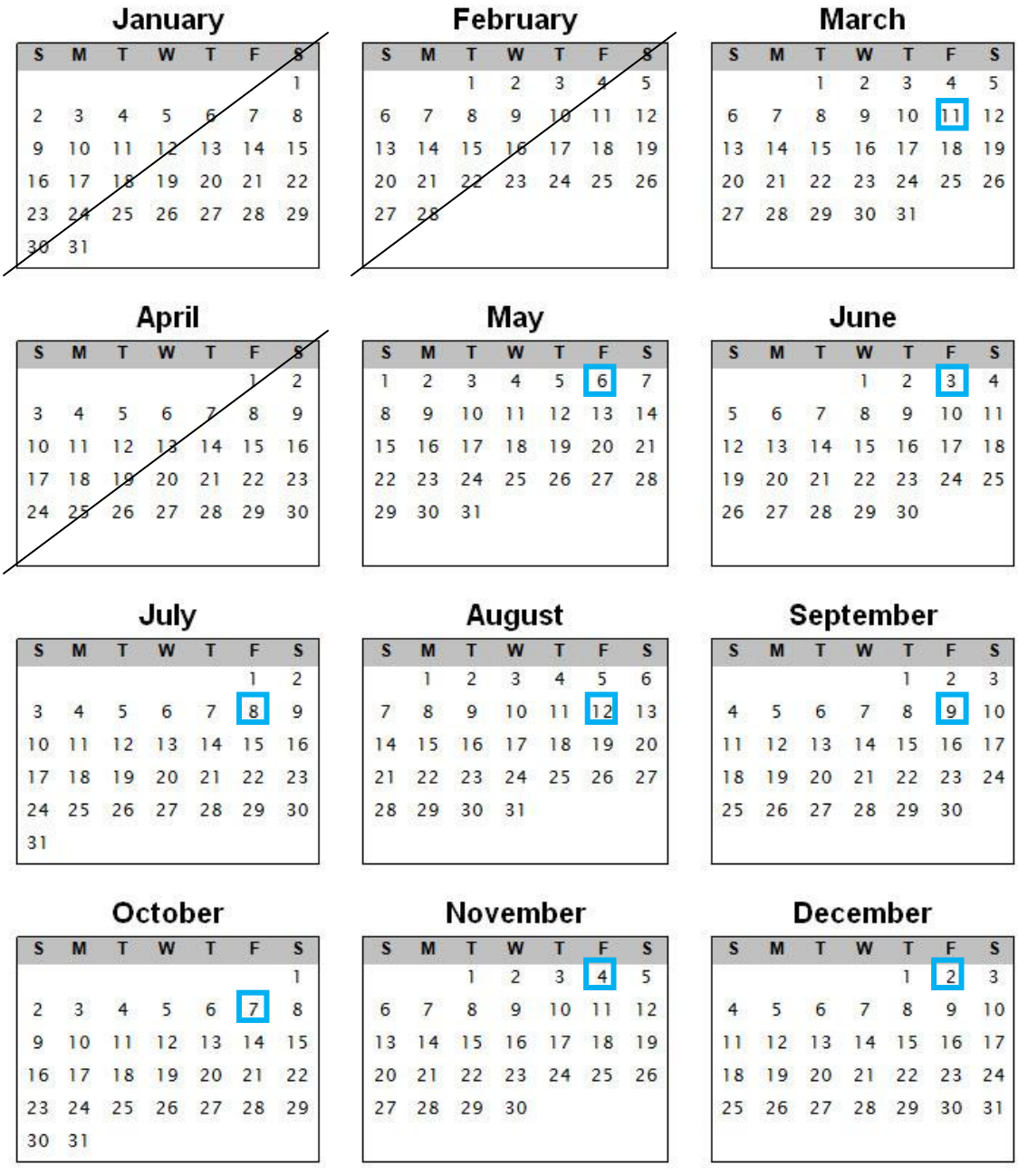
Heide Cassidy, Christina McDougall
P-20 Program Managers
point b Consulting
Heide.Cassidy@ofm.wa.gov, Christina.McDougall@ofm.wa.gov

Connie Michener
Sr IT Consultant
Washington State Information Services Board
connie.michener@dis.wa.gov
360-902-3468

Data Sharing Policy Workgroup Meetings

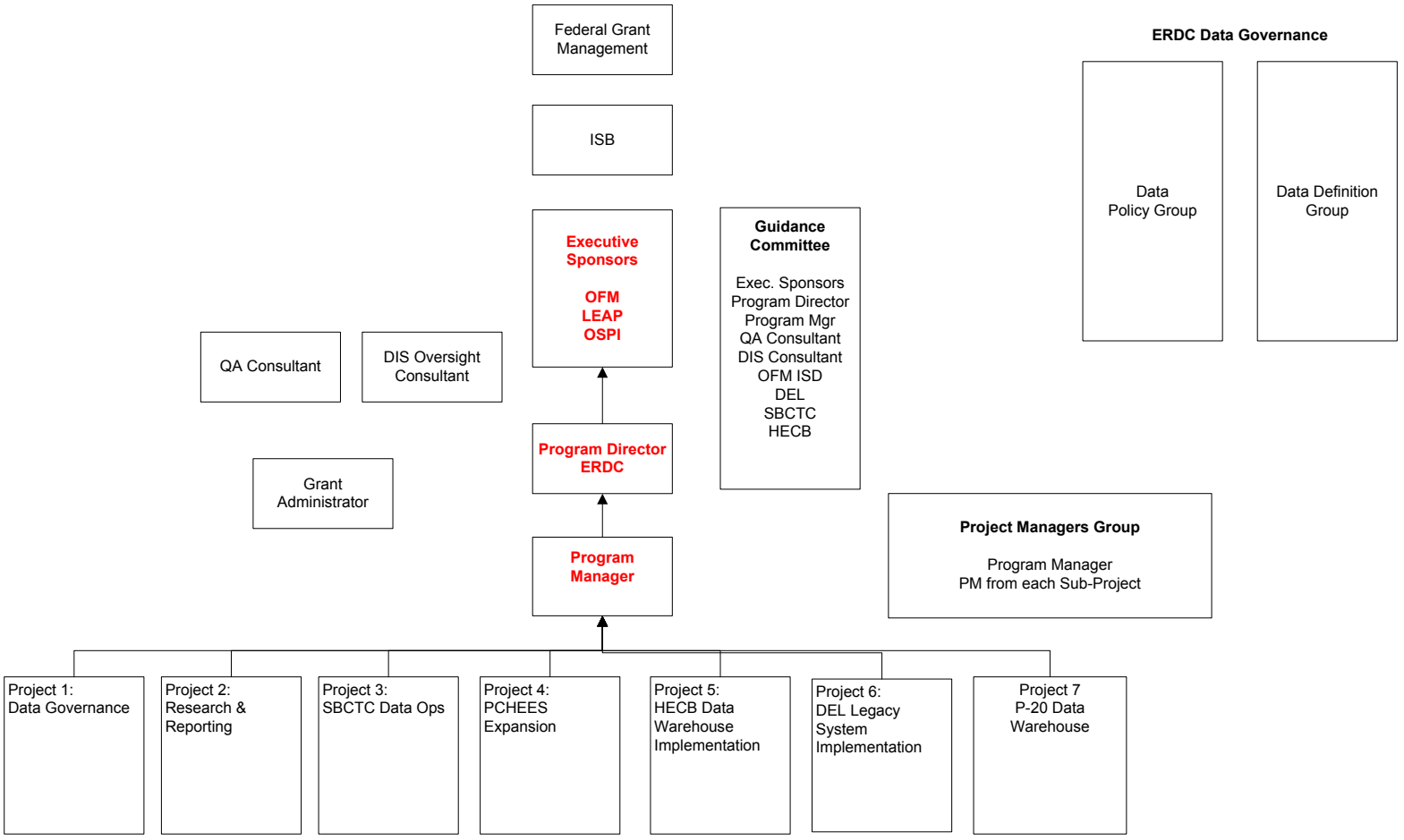
All meetings to be held 8:30 to 10:30 a.m., SBCTC, 1300 Quince St SE, Olympia unless otherwise specified in individual meeting agenda

2011



Rev: 4/1/2011

P-20 Program Governance



P-20 Data Sharing Project Charter

ARRA Grant for Evergreen State P-20 Longitudinal Education Data System

Version 1.1

STATE OF WASHINGTON

Education Research & Data Center
OFFICE OF FINANCIAL MANAGEMENT
FORECASTING DIVISION

P-20 Data Sharing Project Charter

Revision History

Date	Version #	Author	Comments
January 31, 2011	1.0	Deb Came	Original document
February 11, 2011	1.1	Deb Came	Minor revisions following internal review and Executive Sponsor comments

P-20 Data Sharing Project Charter

Document Purpose

The purpose of this document is to define the business objective and outcomes for the project and provide an overview of the scope, high-level schedule and budget, resources required and specific assumptions and constraints under which the project will be completed. This project charter is the authorization for the start of the project and grants the project manager the authority to pursue the project goals and outcomes. Once signed, the charter may be used to guide the general direction of the project throughout its lifecycle.

P-20 Program Background

The Evergreen State P-20 Program, made possible by a $17.3M ARRA grant, is a program focusing on building the foundational capabilities of the State of Washington's Education Research and Data Center (ERDC) to capture, analyze and report on educational longitudinal data ranging from preschool, through K-12, post-secondary education and into the workforce. Deliverables for the three year program include pre-defined research briefs and reports created using existing methods and datasets, the establishment of data governance agreements and processes and a P-20 data warehouse which will be utilized for future reporting projects. Completing these key deliverables will result in significant advancement of ERDC's capacity to meet Washington State's P-20 information needs.

Project Background

"An LDS is not just an information technology (IT) project. *An LDS is a business solution, a way to meet information needs.*" This statement from *Traveling Through Time: The Forum Guide to Longitudinal Data Systems (*page 28*;* http://nces.ed.gov/forum/ldsguide/book1/index.asp) highlights that the purpose of the longitudinal data systems is to produce information. Ultimately, Washington's P-20 longitudinal data system will provide information in a variety of ways that will reach a broad number of stakeholders.

ERDC is bound by the Family Educational Rights and Privacy Act (FERPA) and by existing data-sharing agreements with our partners. ERDC wants to maintain good relationships and preserve the trust of our contributing-data partners. There are current data-sharing agreements that allow ERDC to receive and link data from numerous sectors. While those lay the groundwork, much work remains to determine how data may be shared out of the P-20 system once ERDC has combined individual-level information from multiple sources. Without new agreements, ERDC may not give linked datasets nor report on anything other than highly aggregated analyses.

Vision, Outcomes and Benefits

This project will establish data-sharing and use protocols, confidentiality and privacy requirements, and review processes to assure data owners that data are being used appropriately. The project will also result in a transparent application process for researchers interested in obtaining linked education data. This type of data governance – the agreement of when, how, and what data may be shared – is a crucial component to making the data available to state

agencies and external researchers, particularly when such a large volume of data and numerous contributing agencies are involved.

ERDC's mission is to develop longitudinal information spanning the P-20 system in order to facilitate analyses, provide meaningful reports, collaborate on education research, and share data. This project will expedite that mission by enabling data-sharing with our partner agencies and external researchers. The project vision is to establish a transparent process for sharing data and clear guidelines for data use and reporting.

By opening the door for the sharing of data and thereby enabling additional research and analyses, the project benefits ERDC, the overall SLDS program's creation of a P-20 system, and partner agencies. This project will address critical issues regarding *sharing* the data and making it public. Without this component to the overall grant, there would be a large repository of data but no approved way to make the data available to education researchers and agency staff. It also assures that there will be confidentiality guidelines for data users.

This path will be complementary to the existing K-12 data governance process. The data governance system for K-12 focuses on identifying and prioritizing policy and research questions, data collection, data quality, and managing change in conjunction with the impact on school districts. While the two governance perspectives are related, the P-20 emphasis will be on how to share linked data, including obtaining agreements from data owners, clarifying data use restrictions, and maintaining confidentiality.

Project Scope

In general, this project will determine if ERDC may share data and under what circumstances. It will establish guidelines for restrictions on using and reporting linked education data. It will define a data-request process.

Areas that are *out* of scope for this project:

- Interoperability or other technical aspects of sharing data
- Data element inventory and standard data definitions
- Data quality and accuracy
- "Data Governance" similar to OSPI's data governance group (topics such as gap analyses, identification of research and policy questions, data collection issues)

Major Deliverables or Milestones:

Background information to inform the P-20 system protocols and processes:

(A) Case studies from other states: a compilation of information from approximately five states, outlining their policies and procedures for sharing education data.

(B) Washington State summary: a compilation of information regarding data-sharing policies and processes for Washington P-20/W agencies (DEL, OSPI, SBCTC, HECB, COP, DSHS). This document will provide an outline of existing Washington education agency procedures for sharing individual-level data.

(C) Data Policy Committee meeting (milestone). The P-20 Program will be reaching out to its partners to establish the Program's Data Policy Committee. Once established, ERDC staff will convene the committee and come to a consensus about how group decisions will be made for data-sharing protocols and processes [(D) through (G) below]. The committee will meet regularly to address the protocols incrementally.

Documents that will form P-20 system data-sharing protocols and processes:

(D) Guidelines for sharing individual-level data. This document will outline who is entitled to see what level of data, based on agency affiliation or role. The guidelines will include a matrix (overview) as well as a description of data privileges. It will include broad rules on whether linked data needs to be aggregated, de-identified, or anonymized before it is shared with a given individual or entity.

(E) Reporting guidelines (e.g. cell size restrictions). This document will describe the guidelines for reporting aggregated data. These rules will be designed to minimize the possibility of unintended disclosure of information about an individual.

(F) Review of products. This document will describe the process that categories of users will follow to have their reports reviewed and/or approved.

(G) Criteria and requirements for receiving data. This document will outline the criteria that must be met for someone to receive data. It will address questions such as: Does the data requestor need to meet certain requirements, such as approval through an institutional review board or an online course in maintaining data confidentiality? Will there be other research criteria such as "does the research improve education in WA"? In addition, the document will detail the security protocol for transferring and storing data.

(H) Data-sharing agreement templates. These templates will contain a menu of components that will make up specific data sharing agreements. These templates will be based on the findings and agreements in (D) through (G) above.

(I) Present documents to AAG for review (milestone)

(J) Develop process for data requestors. The documentation will provide data requestors with a step-by-step process to apply for data, including information on how applications are approved. It will answer questions such as expected timeframes for research approval and data receipt and establishment of the structure governing data requests. This includes defining how single-sector data requests will be handled.

(K) Develop approach to prioritize data requests and applications. Given limited resources, ERDC and partner agencies will need a standard approach for prioritizing data requests. This deliverable will develop the prioritization rules as well as any accompanying documentation, such as weighting schemes for data requests.

(L) Post procedures, agreements (templates), policies to website

(M) Pilot data-sharing agreements for DSHS, MESA, and GEAR-UP (HECB). Each of these data-sharing agreements might be done with a collection of bilateral agreements which together will enable the sharing linked data from multiple sources. These will help inform the development of the data-sharing templates (H).

Approach

The P-20 Program will be establishing a Data Policy Committee. The ultimate purpose of the group will be to form agreements about data-sharing protocols and requirements, as outlined in the deliverables listed above. The process for selecting committee members will be validated and the specifics of committee roles and responsibilities will be confirmed. The Data Sharing project is researching and documenting information for this committee to use to draft its foundational policies for the P-20 system.

Specifically, ERDC data-sharing analysts are compiling information about other states' research protocols and agreements for sharing individual-level education data. In addition, staff are summarizing whether Washington education agencies currently share individual-level data and what the guidelines and protocols are for doing so. These two documents will provide context and serve as starting points for discussion about how Washington ought to proceed. The data-sharing project staff will provide background information and make recommendations (as starting points for discussion) to the committee; the committee will make decisions about the final deliverables and policies.

The approach for this project will be to build consensus among agency partners around data-sharing issues. Although there are a number of complex issues to be addressed, they will be taken up incrementally.

Stakeholders

Stakeholder	
Department of Early Learning	Office of Superintendent of Public Instruction
Higher Education Coordinating Board	Council of Presidents
State Board for Community and Technical Colleges	Employment Security Department
Professional Educator Standards Board	Workforce Training and Education Coordinating Board
State Board of Education	Department of Social and Health Services
Mathematics Engineering Science Achievement	K-12 Schools, Districts, and Education Service Districts
Community and Technical Colleges	Public Baccalaureate Institutions

Legislative Evaluation and Accountability Program	Washington State Office of the Attorney General
External researchers (Center for Strengthening Teaching Profession, Center on Reinventing Public Education, College Success Foundation, university researchers, etc.)	

High-Level Schedule

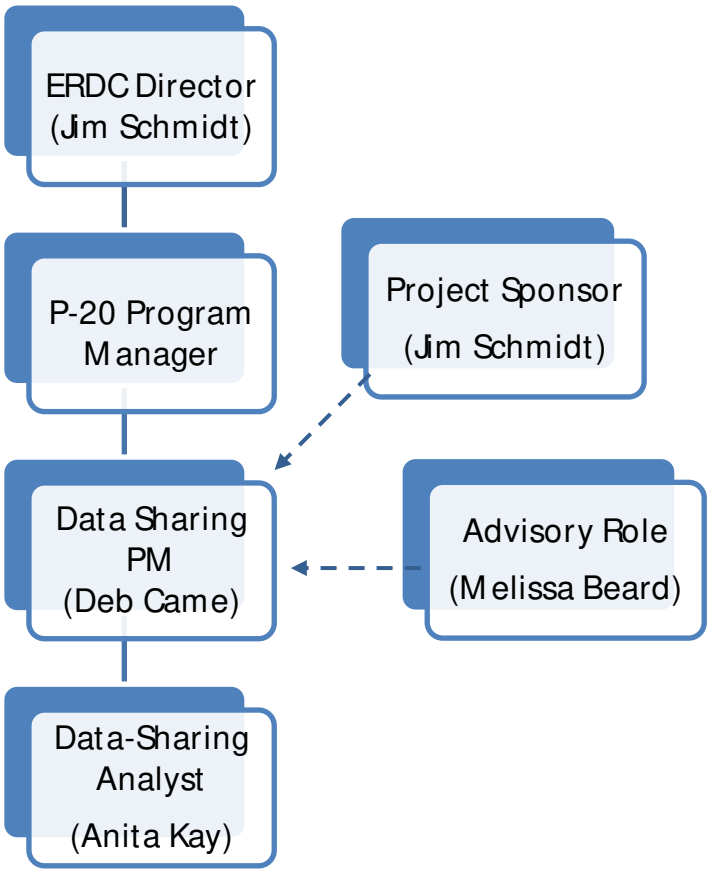
A	Case studies	2/4/2011
B	Summary of WA education agencies' data sharing policies	2/11/2011
C	Kick-off meeting of data policy committee	2/28/2011
D	Sharing guidelines (matrix)	4/30/2011
E	Reporting guidelines (cell size restrictions)	6/30/2011
F	Research product review process	8/31/2011
G	Criteria and requirements for data requestors / recipients	10/31/2011
H	Data sharing agreement templates	1/31/2012
I	AAG review of documents	2/28/2012
J	Develop process for data requestors	5/31/2012
K	Develop approach to prioritize data requests and applications	6/30/2012
L	Post procedures, agreements (templates), policies to website	7/31/2012
M	Pilot agreement: DSHS	6/30/2011
	Pilot agreement: GEAR-UP	4/30/2011
	Pilot agreement: MESA	7/31/2011

P-20 Data Sharing Project Charter

Dependencies on Other P-20 Projects

- Reporting guidelines (E) and sharing guidelines (D) will dictate some specifics for Feedback Reports (Research and Reporting Project). The data-sharing findings will specify who can see what data and to what level of detail (e.g. cell size requirements for reporting). Until those detailed requirements are clarified, the feedback reports can be presented at a highly aggregate level so as not to compromise student identity.
- Pilot data-sharing agreements (M) will need to be signed before research datasets (Research and Reporting Project) can be distributed.
- Sharing guidelines (D) may influence the P-20 data warehouse project, if role-based access is determined to be within the scope of that project.
- Data-sharing agreement templates (H) may depend on data dictionaries (resulting from the P-20 data warehouse project), if individual data items will be listed as a menu in the templates.

Project Organization



Resources Required

The following table describes the high-level responsibilities associated with particular roles or groups for this project under the P-20 Program.

Role	Responsibility and Sourcing
Anita Kay	Data sharing analyst (100% grant; new hire)
Melissa Beard	Data sharing advisor (50%, in-kind)
Deb Came	Data sharing project manager (25%*, in kind)
tbd	Analyst, Council of Presidents (25% FTE; grant)

** 25% of 32 hours per week. Deb works 80% time (32 hours per week).*

Cost/Budget Estimate

See attached.

Assumptions

- Contributing-data agencies will permit data to be shared or are not extremely restrictive about such permissions.
- Contributing-data agencies are able to reach agreement about data-sharing protocols. The schedule assumes, for each deliverable, that the issue can be presented, discussed and a consensus reached over the span of two meetings (and meetings occur approximately monthly).
- Interpretations of legal and other restrictions are not overly complex. For example, FERPA prohibits disclosing information that might allow a "reasonable person" to identify a student. How "reasonable person" is interpreted will affect the complexity of data-sharing protocols and thus may impact the timeline.
- Data requestors / organizations that are pilots to receive P-20 data (GEAR-UP, MESA, and DSHS) will provide data specifics and needs in a timely way such that data-sharing agreements can be developed according to the timeline.

Constraints

- Contributing-data agencies are unwilling to permit data to be shared or are extremely restrictive about such permissions

P-20 Data Sharing Project Charter

High-Level Risks, Impacts and Mitigation Strategies

#	Risk Description	Project Impact	Mitigation Strategy
	Decisions or agreements regarding data-sharing are not made in a timely manner	ERDC would not be able to share research datasets; feedback reports would not be as detailed	Inform data policy group, stakeholders, and steering committee of dates and decision timeframe
	Agency (or agencies) unwilling to have data shared externally (outside ERDC)	ERDC would not be permitted to provide research datasets, or the datasets would be missing key data elements	Clearly outline legal and other issues related to sharing data; Develop assurances regarding data use and confidentiality in the data-sharing policies; Clearly communicate

Charter Signatures

/ s/ Date:

Executive Sponsor – Office of Financial Management,
Assistant Director, Forecasting Division

/ s/ Date:

Project Sponsor – Office of Financial Management,
Director, Education Research & Data Center

/ s/ Date:

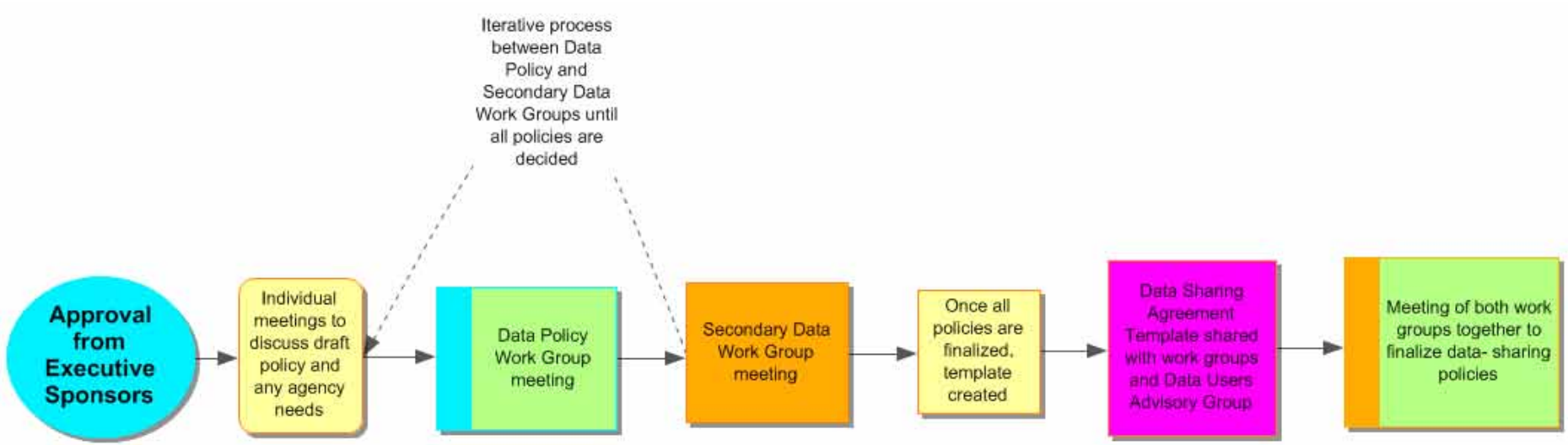
Project Manager – Office of Financial Management,
Education Research & Data Center

cc: Heide Cassidy, Point b, P-20 Program Manager
Christina McDougall, Point b, P-20 Interim Program Manager
Connie Michener, Department of Information Services, Oversight Consultant
Glenn Briskin, Briskin Consulting, External Quality Assurance
Executive Guidance Committee, P-20 Program

PROCESS TO DEVELOP DATA-SHARING POLICIES

Statement of purpose: To meet the needs of ERDC and the grant, ERDC will work with representatives from education and other data contributing agencies to create:

- a data sharing agreement template
- a process to prioritize requests
- a process to notify agencies when their data has been shared
- other products or processes the representatives deem necessary



Deliverables created by this process that form "Washington P-20 Data-Sharing Policies:

4/30	Data Sharing Matrix
6/30	Reporting Guidelines
8/31	Review of Products Guidelines
10/31	Criteria for Receiving Data
1/31/12	Data Sharing Agreement Template
5/31	Data Set Request Process
6/30	Prioritization Process; End of Work Group commitment

Group	Expectations	Members
Data Policy Work Group Main data contributors; most requests will involve at least one, if not all of these data contributors	• Monthly individual meetings • Monthly work group meetings • Discuss materials with appropriate staff prior to individual and monthly meetings • Periodic email or phone call responses when needed • Prepared to approve draft at the meeting	• Department of Early Learning • ECEAP provider • Office of Superintendent of Public Instruction • School district rep • State Board for Community and Technical Colleges • Community or Technical College rep • Higher Education Coordinating Board • Council of Presidents • Public Baccalaureate Institution rep • Employment Security Department • Attorney General's Office-Dave Stolier • Legislative Evaluation and Accountability Program Committee-Tom Jensen (Executive Sponsor) • Office of Superintendent of Public Instruction-Bob Butts (Executive Sponsor)
Secondary Data Work Group Auxiliary data matched to data provided by main data agencies	• Monthly work group meetings • Discuss materials with appropriate staff prior to monthly meetings • Periodic email or phone call responses when needed • Prepared to approve draft at the meeting	• Department of Corrections-admit and education data • Labor and Industries-Apprenticeship data • Workforce Training & Education Coordinating Board-Private Career Colleges data • Department of Retirement Systems-Teacher retirement data
Data Users Advisory Group Provide perspective on process from the consumer side	Provide feedback on draft template and process to prioritize work	• Researchers (Marge Plecki, Dan Goldhaber) • Other agencies (PESB, SBE, DSHS, JLARC, WSIPP) • Kids Count • Gates Foundation • K-12 Groups (WEA, WASA, WSSDA, PSE, AWSP, WA-PTA, WAEYA) • Legislative staff • Others (CCER-Seattle Foundation, New Futures, MESA)

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

SLDS Technical Brief

Guidance for Statewide Longitudinal Data Systems (SLDS)

November 2010, Brief 1
NCES 2011-601

Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records

Contents

SLDS Technical Briefs are intended to provide "best practices" for consideration by states developing Statewide Longitudinal Data Systems.

For more information, contact:
Marilyn Seastrom
National Center for Education Statistics
(202) 502-7303
Marilyn.Seastrom@ed.gov

The National Center for Education Statistics (NCES) is launching a new series of Technical Briefs on various aspects of the protection of personally identifiable information in students' education records. The immediate demand for this work arose from increased federal mandatory reporting (20 U.S.C. § 6311) and the related expansion of record keeping under the Statewide Longitudinal Data Systems (SLDS) (20 U.S.C § 9607; Public Law 111-05 American Recovery and Reinvestment Act of 2009 (ARRA)). This increase in the amount of data published and stored must be balanced against the legal requirements under the Family Educational and Privacy Rights Act (FERPA) to protect personally identifiable information in student education records (20 U.S.C. § 1232g). (Education records include those records that are directly related to a student and are maintained by an educational agency or institution or by a party acting for the agency or institution (34 CFR § 99.3).)

While driven by recent events, the principles and practices that are outlined in this series can be applied more generally to personally identifiable information about students. This series of Technical Briefs is intended to be useful for anyone responsible for the development, maintenance, protection, or use of student record data. This first brief discusses basic concepts and definitions that establish a common set of terms related to the protection of personally identifiable information, especially in education records.

Personally Identifiable Information

The definition of *personally identifiable information* is central to all discussions of privacy and confidentiality. The Office of Management and Budget (OMB) Guidance for the implementation of the Confidential Information Protection and Statistical Efficiency Act of 2002 and OMB Memorandum 07-16, *Safeguarding Against and Responding to the Breach of Personally Identifiable Information,* both state that "The term 'personally identifiable information' refers to information that can be used to distinguish or trace an individual's identity, such as their name, Social Security Number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."

The National Institute of Standards and Technology (NIST) definition in the 2010 publication *Guide to Protecting the Confidentiality of Personally Identifiable Information* (NIST Special Publication 800-122, p. E-1) parallels the OMB definition. Although there is some variation in the wording of the definition across different applications of the term, the OMB definition is the basis for the definition of personally identifiable information across the federal government.

The Family Educational Rights and Privacy Act (FERPA) 2008 regulations (34 CFR § 99) define personally identifiable information for education data and student education records.

Personally identifiable information, as defined in FERPA, includes, but is not limited to:

1. The student's name;
2. The name of the student's parent or other family members;
3. The address of the student or student's family;
4. A personal identifier, such as the student's Social Security Number, student number, or biometric record;
5. Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;
6. Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty;
7. Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates. (34 CFR § 99.3)

Discussions of personally identifiable information frequently use the concepts of *identifiable form* and *direct and indirect identifiers*. The first of these terms was codified in law in the E-Government Act of 2002 (Public Law 107-347). Section 208(d) of that Act states that "In this section, the term 'identifiable form' means any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means" (44 U.S.C. § 3501, note).

The FERPA definition of personally identifiable information calls out specific direct identifiers, such as name, biometric record, Social Security Number, and student number. The FERPA regulations define a biometric record as including measurable biological or behavioral characteristics, such as fingerprints, retina and iris patterns, voiceprints, DNA sequence, facial characteristics, and handwriting (see 34 CFR § 99.3 for full definition).

The FERPA definition refers to "other indirect identifiers such as the student's date of birth, place of birth, and mother's maiden name" and to "other information that, alone or in combination, is linked or linkable to a specific student…" The FERPA definition also includes targeted requests—that is requests where the person requesting the information is trying to get information on a specific student. For example, if there was a rumor published in the local paper that a public official was disciplined for cheating during his senior year in high school, a request to the high school for the disciplinary records of students who were caught cheating during the year the public official was a senior would be considered a targeted request.

OMB Memorandum M-03-22 *OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002* and OMB Memorandum M-07-16 *Safeguarding Against and Responding to the Breach of Personally Identifiable Information* provide additional examples of direct and indirect identifiers. Direct identifiers include information that relates specifically to an individual such as the individual's residence, including for example, name, address, Social Security Number or other identifying number or code, telephone number, e-mail address, or biometric record. *Indirect identifiers* include information that can be combined with other information to identify specific individuals, including, for example, a combination of gender, birth date, geographic indicator, and other descriptors.

The 2010 NIST guide extends the list of examples of indirect identifiers to include place of birth, race, religion, weight, activities, employment information, medical information, education information, and financial information (NIST 2010 Special Publication 800-122, p. 2-2).

FERPA allows the public release of some personally identifiable student information as school directory information (20 U.S.C. § 1232g

(b)(1)), where *directory information* is defined in the 2008 FERPA regulations as "information contained in an education record of a student that would not generally be considered harmful or an invasion of privacy if disclosed" (34 CFR § 99.3). The FERPA regulations also specify that directory information may not include a student's Social Security Number or an identification number that is used to access the student's education record. The FERPA regulations require that educational agencies or institutions provide public notice to parents of students or eligible students of the types of personally identifiable information that are designated as directory information (34 CFR § 99.37). The parent or the eligible student must be given the right to refuse to have any or all of the student's information released as directory information. (An "eligible student" is a student who has reached 18 years of age or is attending a postsecondary institution (34 CFR § 99.3).)

The 2008 FERPA regulations state that "Directory information includes, but is not limited to, the student's name; address; telephone listing; electronic mail address; photograph; date and place of birth; major field of study; grade level; enrollment status (e.g., undergraduate or graduate, full-time or part-time); dates of attendance; participation in officially recognized activities and sports; weight and height of members of athletic teams; degrees, honors and awards received; and the most recent educational agency or institution attended" (34 CFR § 99.3).

In the Notice of Proposed Rule Making (NPRM) for the 2008 FERPA regulations, the U.S. Department of Education recognized that the risk of identifying a student in aggregate data is cumulative and related to previous releases of data from student education records in both directory information and aggregate reports that are assumed to protect personally identifiable information, as well as to data from external sources. Furthermore, in acknowledging that these risks have increased as a result of new technologies and methods that emerged since FERPA was enacted in 1974, the Department advised "that parties should minimize information released in directories to the extent possible" (73 Fed. Reg. 15574-602, March 24, 2008).

Privacy and Confidentiality

The terms privacy and confidentiality are often invoked in discussions about rights and responsibilities when it comes to student records; in fact, they are often used interchangeably even though they have distinct meanings. So exactly what does each of these terms mean?

Privacy Defined

The concept of *privacy* relates to individual autonomy and each person's control over their own information (Report of the National Academy of Science 1993 Panel Report *Private Lives and Public Policies,* p. 3). This includes each person's right to decide when and whether to share personal information, how much information to share, and the circumstances under which that information can be shared (Report of the National Academy of Science 1993 Panel Report *Private Lives and Public Policies,* p. 22).

The 2009 National Academy of Sciences Report from the Committee on National Statistics and the Center for Education Workshop, *Protecting Student Privacy and Facilitating Education Research,* defined privacy as "…an individual's control over who has access to information about him or her. The concept of privacy is relevant to what personal information becomes data" (Summary of the Committee on National Statistics' 2009 Workshop on *Protecting Student Records and Facilitating Education Research,* p. 3).

In the context of student education records and FERPA, privacy pertains to the rights of parents and eligible students to inspect and review the students' education records, to seek to amend education records, to consent to the release of personally identifiable information from education records for any disclosures that are not authorized in law, and to refuse to have personally identifiable information that is designated as directory information publicly released (20 U.S.C. § 1232g, 34 CFR §§ 99.7, 99.37).

Confidentiality Defined

Confidentiality relates to the management of another individual's personally identifiable information. In a 2009 National Academy of Sciences, Institute of Medicine report, confidentiality is defined as referring to the obligations of those who receive personal information about an individual to respect the individual's privacy by safeguarding the information (Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule, 2009, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research,* pp. 17–18).

These concepts are echoed in the 2009 National Academies of Sciences workshop report on protecting student records in which confidentiality is defined as "protection against the release of personal information. An important distinction is that privacy pertains to individuals; confidentiality to their information" (Report of the National Academy of Science 2009 Workshop *Protecting Student Privacy and Facilitating Education Research,* p. 4).

Legal and ethical responsibilities to protect against the release of personal information must be respected and enforced even if some of the same information is already in the public domain. The fact that some of the information is already in the public domain can make use or disclosure of other information more sensitive.

Disclosures of Confidential Information

These definitions introduce the concept of protecting personally identifiable information from release. This is also referred to as protecting personally identifiable information from *disclosure*. Under FERPA, "Disclosure means to permit access to or release, transfer, or other communication of personally identifiable information contained in education records by any means including oral, written, or electronic means, to any party except the party identified or the party that provided or created the record" (34 CFR § 99.3).

There are three types of disclosure—*authorized, unauthorized, and inadvertent*. FERPA *authorizes* or permits specific users and uses of personally identifiable information in student education records without the written consent of the parent or eligible student. These authorized disclosures include, but are not limited to, the following:

- other school officials, including teachers within the agency or institution who have legitimate educational interests;
- officials of another school, school system or postsecondary institution in which the student seeks to enroll;
- authorized representatives of the Comptroller General of the United States, Attorney General of the United States, the Secretary of the U.S. Department of Education, and state and local educational authorities;
- in connection with financial aid for which a student has applied or received;
- State and local officials or authorities to whom access is granted under state statute;
- organizations conducting studies for, or on behalf of, educational agencies or institutions for the purpose of developing, validating, or administering predictive tests, administering student aid programs, and improving instruction, subject to confidentiality and privacy conditions (including a written agreement);
- accrediting organizations for accrediting purposes;
- parents of a dependent student;
- information designated as directory information;
- a parent of a student who is under age 18 and not enrolled in postsecondary education;
- a student who has reached age 18 or enrolled in postsecondary education;
- in connection with a health or safety emergency (see 34 CFR § 99.31 for additional details and exceptions).

An *unauthorized* disclosure occurs when personally identifiable information from a student's education record is made available to a third party who does not have legal authority to access the information. An *inadvertent* disclosure occurs when information about an individual is unintentionally revealed through

information released to the public. This might happen, for example, through a security breach of the electronic system that is used to maintain and access the education records, as a result of a teacher or administrator leaving paper reports that include personally identifiable information in an unsecured location, or as a result of identifiable information about a student that can be derived from published summary statistics that were not fully protected.

The National Academy workshop report on protecting student privacy makes a further distinction between the confidential information in a student record that includes personal information and statistical reports derived from that information. The report cites as an example the fact that while a parent has the right to control information pertaining to the fact that his or her child is enrolled in a specific school, a summary statistic of the number of students enrolled in a school does not violate confidentiality and thus does not constitute a disclosure. In other words, it is not a disclosure or a violation of the confidentiality of the information in the data when personal information for a number of students is combined to produce a statistical report (Report of the National Academy of Science 2009 Workshop *Protecting Student Privacy and Facilitating Education Research,* p. 4).

While this is true in the case of a summary enrollment count, it is important to understand that even with statistical reports care must be taken to avoid *inadvertent* disclosures. Disclosures of this type are unintentional and occur when data in a student level file or aggregate data in tabulations allow the data user to identify a student, known as *identity disclosure,* or when data in a student level file or aggregate data in tabulations reveals sensitive information about a student, known as an *attribute disclosure.* For example, a statistical report of student assessment results for Hispanic third-graders in a specific school shows that there were students in this subgroup who scored in each of four different achievement levels. Knowing that these students were distributed across the four achievement levels does not reveal or disclose any information about an individual Hispanic third-grader's performance in that school. However, a statistical report for a different school shows that all of the Hispanic third-graders scored below the target performance level of proficient, and all of the White third-graders scored at or above the proficient level. The report for the second school reveals or discloses information about the performance of both White and Hispanic third-graders in this school—specifically, that each of the White third-graders reached or exceeded the performance target, while each of the Hispanic third-graders in the school failed to reach the target performance level. This release results in an attribute disclosure since specific performance can be associated with all of the students in clearly definable subgroups. (Preventing this type of inadvertent disclosure is the focus of a companion SLDS Technical Brief, *Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting.*)

Protecting Confidentiality Through De-Identification and Anonymization of Data

Other terms that are used in discussing confidentiality include de-identification and anonymization. These concepts are central to protecting against disclosures in data files that are shared with external education researchers. The term *de-identified* information is used to describe records that have enough personally identifiable information removed or obscured such that the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual. The FERPA 2008 regulations subsection on de-identified records allows for the nonconsensual release of student level information from education records, provided that (1) all personally identifiable information is removed and (2) there is a reasonable determination that a student's identity is not personally identifiable. In making this determination, both single and multiple data releases from the education records should be taken into account along with other information available from other sources (34 CFR § 99.31(b)(1)).

In the 2008 issuance of the Final Rule for revisions to the FERPA regulations, the Department of Education referred interested parties to the Federal Committee on Statistical Methodology's Statistical Policy Working Paper 22 *Report on Statistical Disclosure Limitation Methodology* for advice on ways to de-identify student level data (73 Fed. Reg. 74806-35, Dec 9, 2008). The Working Paper includes techniques that can be used to protect against disclosures in student level records in a data file as well as techniques that can be used to protect against disclosures in aggregate tabular reports.

Techniques described that can be used to protect student level data include generalizing the data by grouping continuous values and applying top and bottom coding to either continuous or categorical data to avoid outliers; suppressing the data by deleting entire records or parts of records; introducing "noise" into the data by adding small amounts of variation into selected data; swapping the data by exchanging certain data elements in one record with the same data elements from a similar record; blanking and imputing for randomly selected records; and blurring by replacing data with the average value by replacing a selected value (e.g., an outlier) of a data element with the average value for that data element for the entire group.

Techniques described to avoid disclosures in aggregate tabular data include establishing minimum cell sizes, suppression, complementary suppression, random rounding, controlled rounding, controlled tabular adjustment, and special rules to protect against disclosures that might include additional restrictions on publishing such as requiring results on more than one cell in a distribution, requiring certain size categories, and collapsing across categories.

Once a data file is de-identified, the FERPA regulations indicate that a re-identification code may be attached to the data file so that the file can be released for use for education research (34 CFR § 99.31(b)(2)). While the de-identified data file with a re-identification code does not provide external researchers with personally identifiable information about students, a researcher is able to return to the source that issued the data to request additional data elements that can be added using the re-identification code.

The re-identification code should be independent of any of the personally identifiable information. Only a limited number of staff should have knowledge of the method used to produce the code. Under FERPA, the re-identification code (1) may not be used for any purpose other than matching the de-identified records to the source to obtain additional information for education research; (2) may not be used to identify a student or personally identifiable information about a student; and (3) may not be based on a student's Social Security Number or other personal information (34 CFR § 99.31(b)(2)).

To understand how this would work, take the case of a school district that received a data request from an external researcher who is interested in analyzing academic gains for students who participated in an afterschool enrichment program. To do this, the district creates a fully de-identified data file that includes the relevant individual student records drawn over the researcher-specified time period for a subset of data elements that do not identify individual students. During the course of the analysis, the researcher discovers that several additional data elements and an additional year of data are needed to produce a robust analysis. The district data manager uses the re-identification codes to identify the same set of students and create an extract file that includes the additional data elements and an additional year of data for those students. The researcher uses the code to link the new data to the existing analysis file and proceeds with the analysis.

Anonymization takes the data one step beyond de-identification. That is, anonymized data are data that have been de-identified, *and* they do not include a re-identification code. In an anonymized data file, the student case numbers in the data records cannot be linked back to the original student record system. Returning to the examples discussed above, anonymized data would not be useful to staff using data to monitor the progress and performance of individual students. However, if a professor at a university reads the research report from the analysis of academic gains of students in the afterschool enrichment program and decides that he or she would like to have a class of graduate students apply different analytic procedures to see if the results can be replicated, an anonymized file could be produced from the de-identified file used by the researchers to serve this purpose. To do this, the re-identification code must be removed and the file should be reviewed to ensure that additional statistical disclosure techniques do not need to be applied. The documentation for the anonymized data file should identify any disclosure limitation techniques that were applied and their implications for the analysis.

Data Stewardship and Privacy Framework

Maintaining personally identifiable information in student education records carries both legal and ethical responsibilities for protecting the information and for ensuring the proper handling and use of the information. These concepts are part of data stewardship. The American Statistical Association's Committee on Privacy and Confidentiality cites the U.S. Census Bureau definition of data stewardship as an "organizational commitment to ensure that identifiable information is collected, maintained, used, and disseminated in a way that respects privacy, ensures confidentiality and security, reduces reporting burden, and promotes access to statistical data for public policy."

These elements of data stewardship are enacted in the various federal privacy and confidentiality laws that govern the use of personally identifiable information, including the Privacy Act of 1974, the Paper Work Reduction Act of 1980, the FERPA of 1974, the Education Sciences Reform Act of 2002 (and related authorizing laws from 1988 and 1994), and the Confidential Information Protection and Statistical Efficiency Act of 2002. They are also included in a set of tenets known as Privacy Principles or, alternatively, as Fair Information Practices. These Privacy Principles, which have been credited as forming the framework for most modern privacy laws, can be traced to the 1973 U.S. Department of Health, Education, and Welfare (HEW) report *Records, Computers and the Rights of Citizens,* Report of the Secretary's Advisory committee on Automated Personal Data Systems. The 1973 report recommended the enactment of a *Federal Code of Fair Information Practice,* consisting of a set of privacy principles.

The 1973 privacy principles set the stage for the passage of three landmark pieces of legislation—the Privacy Act, FERPA, and the Paper Work Reduction Act. Parental complaints about intrusive surveys and other data-collection activities have been cited as one reason for the enactment of FERPA (The 1977 Privacy Protection Commission, Chapter 10 Record Keeping in the Education Relationship).

These privacy principles were described in the 1973 HEW Report as safeguard requirements for data systems that include personally identifiable information. Each of these principles can be found in FERPA and the FERPA Regulations. The first principle, that there should be no secret records of personal data, is evident in the required FERPA annual notification to parents and eligible students of their right to inspect and review the student's education records (20 U.S.C. § 1232g (e); 34 CFR § 99.7). The second principle, that an individual has the right to know what personal information is retained and how it is used, is operationalized through the right to inspect and review the student's education record (20 U.S.C. § 1232g (e); 34 CFR § 99.10) and through the permissible uses of the information which are described in 20 U.S.C. § 1232g (b) and 34 CFR § 99.31. The third principle, the limitation of alternative uses of personal information without consent, is evident in the FERPA requirement that the parent or eligible student provide written consent for the student's information to be used for any purpose not specified in law (20 U.S.C. § 1232g (b)(1); 34 CFR § 99.30). The fourth principle, that an individual has the right to correct or amend a record of personal information, is addressed in law through the requirement that the parent or

The 1973 Fair Information Practices included five principles:

1. There must be no personal data record keeping systems whose very existence is secret.
2. There must be a way for an individual to find out what information about him or her is in a record and how it is used.
3. There must be a way for an individual to prevent information about him that was obtained for one use from being used or made available for other purposes without his consent.
4. There must be a way for an individual to correct or amend a record of identifiable information about him.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for the intended use and must prevent misuse of the data.

an eligible student be provided an opportunity to challenge and seek a correction of the content of a student's record "to insure that the records are not inaccurate, misleading, or otherwise in violation of the student's privacy" (20 U.S.C. § 1232g (a)(2); 34 CFR § 99.20). Finally, the fifth principle, the obligation to prevent the misuse of any personal data maintained and to ensure the reliability of the data for the intended use, is codified in the limitations on permissible uses of the information (20 U.S.C. § 1232g (b)(1)(A-J); 34 CFR §§ 99.31, 99.33–99.35).

The Privacy Act of 1974 called for a Commission to evaluate the implementation of the Privacy Act. The resulting Privacy Protection Study Commission expanded the HEW Commission's list of five Fair Information Practices to a set of eight principles, variations of which have been adopted broadly nationally and internationally.

The recent **Department of Homeland Security and Chief Information Officer Fair Information Practice Principles** include the following:

1. **TRANSPARENCY—**providing notice to the individual regarding the collection, use, dissemination, and maintenance of personally identifiable information.

2. **INDIVIDUAL PARTICIPATION AND REDRESS—**involving the individual in the process of using personally identifiable information and seeking individual consent for the collection, use, dissemination, and maintenance of personally identifiable information. Providing mechanisms for appropriate access, correction, and redress regarding the use of personally identifiable information.

3. **PURPOSE SPECIFICATION—**specifically articulating the authority that permits the collection of personally identifiable information and specifically articulating the purpose or purposes for which the personally identifiable information is intended to be used.

4. **DATA MINIMIZATION AND RETENTION—**only collecting personally identifiable information that is directly relevant and necessary to accomplish the specified purpose(s). Only retaining personally identifiable information for as long as is necessary to fulfill the specified purpose(s).

5. **USE LIMITATION—**using personally identifiable information solely for the purpose(s) specified in the public notice. Sharing information should be for a purpose compatible with the purpose for which the information was collected.

6. **DATA QUALITY AND INTEGRITY—**ensuring, to the greatest extent possible, that personally identifiable information is accurate, relevant, timely, and complete for the purposes for which it is to be used, as identified in the public notice.

7. **SECURITY—**protecting personally identifiable information (in all media) through appropriate administrative, technical, and physical security safeguards against risks such as loss, unauthorized access or use, destruction, modification, or unintended or inappropriate disclosure.

8. **ACCOUNTING AND AUDITING—**providing accountability for compliance with all applicable privacy protection requirements. Including all identified authorities and established policies and procedures that govern the collection, use, dissemination, and maintenance of personally identifiable information. Auditing for the actual use of personally identifiable information to demonstrate compliance with established privacy controls.

SLDS Technical Briefs on Privacy

The principles of the Fair Information Practice Principles provide the framework for a sound privacy and confidentiality data protection program. The information practices are recurring themes in each of the reports in the NCES series of Technical Briefs on various aspects of the protection of personally identifiable information in students' education records. In addition to this Technical Brief, the Briefs include the following topics.

Data Stewardship

An understanding of the principles of data stewardship at the school, district, and state levels provides an essential foundation for ensuring student privacy. Data stewardship starts with decisions as to what personally identifiable information is needed to successfully monitor each student's progress through the education system. Data stewardship also involves a commitment to ensuring that personally identifiable information is collected, maintained, used, and disseminated in a way that respects privacy, ensures confidentiality and security, and promotes access to data for policy formation. To provide data stewardship, there is a need for clearly established policies and procedures that govern collection, storage, processing, and access to an individual student's education records. Role-based or managed access to individual data is one key element of data stewardship. Specifically, policies and procedures should identify who within an educational agency or institution is authorized to access the records and the conditions under which they may be accessed and released. Policies could cover topics such as limiting access to "need to know" and rules and procedures to prohibit authorized users from looking at information they are not authorized to access (i.e., browsing). Procedures could include the use of signed statements of nondisclosure for authorized staff, specified methods for access and retrieval of individual records, and the identification of a secure location for their use (Fair Information Practice Principles 1 through 8).

Electronic Data Security

The development and maintenance of an efficient state longitudinal data system requires the use of an electronic record system. Because these data systems include personally identifiable student information, they should be in an electronically secure environment. All data and the hardware, software, and network infrastructure should be firewall secure and password protected to be safe from unauthorized external access. Furthermore, electronic encryption or secure networks should be used to transmit data with personally identifiable information between different entities (e.g., between the district and the state agencies) (Fair Information Principles 7 and 8).

Statistical Methods for Data Protection in Aggregate Reporting

Using information contained in student education records and related state longitudinal data systems for reporting and research requires reporting information on aggregates of students. Such reporting requires the identification and use of appropriate disclosure avoidance techniques to protect the identity of individual students in publicly available information. Because education data are reported at multiple levels (i.e., school, district, state, and federal) and in external studies, care must be taken to avoid inadvertent disclosures that can occur through comparisons of released data across reporting levels. Some current practices can result in inadvertent disclosures. A set of reporting rules that offers one approach to protecting identifiable student information in aggregate reported data will be presented. The goal of these reporting rules is to have an easy to understand and implement set of steps that can be used to protect personally identifiable student data in aggregate data. To facilitate the implementation of these rules, NCES will also provide an online tool that can be used to implement the rules (Fair Information Principle 7).

External Data Use and Written Agreements

The FERPA regulations include provisions that permit the nonconsensual release of de-identified data sets with re-identification codes to facilitate external research (34 CFR § 99.31(b)). Each of these concepts is discussed. In addition, the FERPA regulations include provisions that permit state and local educational authorities to redisclose personally identifiable information from education records to organizations conducting studies pursuant to the terms of

34 C.F.R. § 99.31(a)(6)(ii)(C) through the use of written agreements that identify and codify the terms of data sharing. Recommended components of these agreements will be discussed and a model template for an agreement will be provided (Fair Information Principles 5, 6, and 7).

Training

To successfully implement a privacy and confidentiality program for student education records, the managers of student education record systems should provide relevant staff at the state, district and school levels with periodic training to inform them of the continuing data use and data protection provisions in FERPA and other applicable privacy and security statutes and to train them on methods for compliance. These training needs are identified, with suggestions for specific content, in the guidance documents.

Data stewards and analysts will need training on newly identified disclosure limitation procedures and reporting rules for the increased protection of personally identifiable information in student education records. The technology staff should be trained on secure data transmissions, and data stewards and data managers should be trained on internal access rules and procedures and on the use of written data agreements and signed statements of nondisclosure (All Fair Information Principles, but especially 5, 6, 7, and 8)

Summary

This series of SLDS Technical Briefs is intended to open a conversation with education practitioners responsible for developing and using electronic student record systems about student privacy considerations that arise in these record systems.

NCES welcomes input on this and each of the forthcoming SLDS Technical Briefs on Privacy. You may direct comments to SLDStechbrief@ed.gov.

References

American Statistical Association, Committee on Privacy and Confidentiality. *Key Terms/ Definitions in Privacy and Confidentiality.* Alexandria, VA: Retrieved from http://www.amstat.org/committees/pc/keyterms.html on 6/17/2010.

Code of Federal Regulations, Title 34—Education, Part 99. *Family Educational and Privacy Rights,* (34CFR99). Washington, DC: GPO Access e-CFR. Retrieved from http://ecfr.gpoaccess.gov/cgi/t/text/ext-idx?c=ecfr&sid=44d350c26fb9cba4a156bf805f297c9e&tpl=/ecfrbrowse/Title34/34cfr99_main_02.tpl on 9/9/2010.

Duncan, George T., Jabine, Thomas B. and de Wolf, Virginia A., Editors (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Panel on Confidentiality and Data Access, National Research Council. Washington, DC: National Academy Press.

Federal Register, Office of Management and Budget. *Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002* (CIPSEA). Washington, DC: Vol. 72, No. 115/ Friday, June 15, 2007. Retrieved from http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2007/061507_cipsea_guidance.pdf on 9/9/2010.

Federal Register, Part II Department of Education. *Family Educational Rights and Privacy; Proposed Rule (34 CFR Part 99).* Washington, DC: Vol. 73, No. 57/ Monday, March 24, 2008.

Federal Register, Part II Department of Education. *Family Educational Rights and Privacy; Final Rule (34 CFR Part 99).* Washington, DC: Vol. 73, No. 237/ Tuesday, December 9, 2008.

McCallister, E., Grance, T., and Scarfone, K. (2010). *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII): Recommendations of the National Institute of Standards and Technology* (NIST Special Publication 800-122). National Institute of Standards and Technology, U.S. Department of Commerce. Washington, DC: Retrieved from http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf on 5/4/2010.

National Academy of Sciences, Committee on National Statistics and the Center for Education, Workshop (2009). *Protecting Student Privacy and Facilitating Education Research.* Washington, DC: National Academy Press.

National Academy of Sciences, Institute of Medicine, Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule (2009). *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research.* Washington, DC: National Academy Press.

Office of Management and Budget, Federal Committee on Statistical Methodology, (2005). Statistical Policy Working Paper 22, *Report on Statistical Disclosure Limitation Methodology.* Retrieved from http://www.fcsm.gov/working-papers/spwp22.html on 9/9/2010.

Office of Management and Budget. OMB Memorandum M-03-22, *OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002.* Retrieved from http://www.whitehouse.gov/omb/memoranda_m03-22/ on 9/9/2010.

Office of Management and Budget. OMB Memorandum M-07-16, *Safeguarding Against and Responding to the Breach of Personally Identifiable Information.* Retrieved from http://www.whitehouse.gov/sites/default/files/omb/assets/omb/memoranda/fy2007/m07-16.pdf on 9/9/2010.

Public Law 107-347, E-Government Act of 2002, Title II, Sec. 208 (d). *Privacy Provisions.* Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ347.107.

Public Law 111-05, American Recovery and Reinvestment Act of 2009, Title VIII—Departments of Labor, Health and Human Services, and Education, and Related Agencies, Institute of Education sciences, Stat. 184, Washington DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_public_laws&docid=f:publ005.111 on 9/9/2010.

U.S. Code, Title 5—Government Organization and Employees, Chapter 5—Administrative Procedure, Subchapter II—Administrative Procedure, Section 552a. *Records Maintained on Individuals (Privacy Act),* (5USC522a). Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/usc.cgi?ACTION=RETRIEVE&FILE=$$xa$$busc5.wais&start=312761&SIZE=77292&TYPE=TEXT on 9/9/2010.

U.S. Code, Title 20—Education, Chapter 31—General Provisions Concerning Education, Subchapter III—General Requirements and Conditions Concerning Operation and Administration of Education Programs: General Authority of Secretary, Part 4—Records, Privacy, Limitation on Withholding Federal funds, Section 1232g. *Family Educational and Privacy Rights,* (20USC1232g). Washington, DC: GPO Access. Retrieved from http://frwebgate4.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=799486197532+0+1+0&WAISaction=retrieve on 9/9/2010.

U.S. Code, Title 44—Public Printing and Documents, Chapter 35—Coordination of Federal Information Policy, Subchapter I—Federal Information Policy, Section 3512. *Public Protection (Paperwork Reduction Act).* (44USC3512). Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/usc.cgi?ACTION=RETRIEVE&FILE=$$xa$$busc44.wais&start=978642&SIZE=2355&TYPE=TEXT on 9/9/2010.

U.S. Code, Title 20—Education, Chapter 70—Strengthening and Improvement of Elementary and Secondary Schools, Subchapter I—Improving the Academic Achievement of the Disadvantaged, Part A—Improving Basic Programs Operated by Local Educational Agencies, Subpart 1—Basic Program Requirements, Section 6311. *State Plans,* (20USC6311). Washington, DC: GPO Access. Retrieved from http://frwebgate2.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=bULwJH/21/1/0&WAISaction=retrieve on 9/9/2010.

U.S. Code, Title 20—Education, Chapter 76—Education Research, Statistics, Evaluation, Information, and Dissemination, Subchapter I—Education Sciences Reform, Section 9573. *Confidentiality,* (20USC9573). Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/usc.cgi?ACTION=RETRIEVE&FILE=$$xa$$busc20.wais&start=10326022&SIZE=10154&TYPE=TEXT on 9/9/2010.

U.S. Code, Title 20—Education, Chapter 76—Education Research, Statistics, Evaluation, Information, and Dissemination, Subchapter II—Educational Technical Assistance, Section 9607. *Grant Program for Statewide, Longitudinal Data Systems,* (20USC9607). Washington, DC: GPO Access. Retrieved from http://frwebgate3.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=FKr6BA/0/1/0&WAISaction=retrieve on 9/9/2010.

U.S. Department of Health and Human Services, Report of the HEW Secretary's Advisory Committee on Automated Personal Data Systems (1973). *Records, Computers and the Rights of Citizens.* Washington, DC: Retrieved from http://aspe.hhs.gov/datacncl/1973privacy/tocprefacemembers.htm on 5/11/2010.

U.S. Privacy Protection Study Commission (1977). *Personal Privacy in an Information Society.* Retrieved from http://epic.org/privacy/ppsc1977report/c1.htm on 9/9/2010.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

SLDS Technical Brief

Guidance for Statewide Longitudinal Data Systems (SLDS)

November 2010, Brief 2
NCES 2011-602

Data Stewardship: Managing Personally Identifiable Information in Electronic Student Education Records

Contents

SLDS Technical Briefs are intended to provide "best practices" for consideration by states developing Statewide Longitudinal Data Systems.

For more information, contact:
Marilyn Seastrom
National Center for Education Statistics
(202) 502-7303
Marilyn.Seastrom@ed.gov

The growth of electronic student data in America's education system has focused attention on the ways these data are collected, processed, stored, and used. The use of records in Statewide Longitudinal Data Systems to follow the progress of individual students over time requires maintaining student education records that include information that identifies individual students. The sensitivity of some of the personally identifiable information in student records increases the level of concern over these data. Administrators and data managers can help ensure the protection of personally identifiable information in the student records they maintain by developing and implementing a privacy and data protection program. The principles embodied in the Fair Information Practices adopted in the United States by the Federal Chief Information Officers Council and the Department of Homeland Security, coupled with the Family Educational Rights and Privacy Act (FERPA) and related regulations, provide a foundation for such a program.

Data Stewardship Defined

In 1973, the Department of Health Education and Welfare (HEW) report *Records, Computers and the Rights of Citizens: Report of the Secretary's Advisory Committee on Automated Personal Data Systems* discussed the need to "maintain data in the system with such accuracy, completeness, timeliness, and pertinence as is necessary to assure accuracy and fairness in any determination relating to an individual's qualifications, character, rights, opportunities, or benefits that may be made on the basis of such data" (pg. 6, Chapter IV). This was codified in the Privacy Act of 1974 (5 U.S.C. § 552a(g)(1)(C). More recently, on their website, the American Statistical Association's Committee on Privacy and Confidentiality cites the Census Bureau's definition of data stewardship as an "organizational commitment to ensure that identifiable information is collected, maintained, used, and disseminated in a way that respects privacy, ensures confidentiality and security, reduces reporting burden, and promotes access to statistical data for public policy." These two sets of requirements can be combined and tailored to education data as follows:

> Data stewardship is an organizational commitment to ensure that data in education records, including personally identifiable information:
>
> » Are accurate, complete, timely, and relevant for the intended purpose;
>
> » Are collected, maintained, used, and disseminated in a way that respects privacy and ensures confidentiality and security;
>
> » Meet the goals of promoting access to the data for evaluating and monitoring educational progress and educational programs; and
>
> » Meet the goals of assuring accuracy to ensure that decisions relating to an individual student's rights and educational opportunities are based on the best possible information.

These requirements are best operationalized through written policies and procedures. Typically, in a system with multiple uses and users, the task of establishing and promulgating policies and procedures is assigned to a Governance Committee that includes representatives of management, legal counsel, the data system administrator, data providers, data managers, and data users. The members representing these different stakeholders should be appointed to the Governance Committee by the head of the state education office, school district, or school, depending on the level where the affected data are held. This group should be established to work collaboratively to develop the policies and procedures for a privacy and data protection program. These policies would then be implemented by the data system administrator through the ongoing management of data collection, processing, storage, maintenance, and use of student records. Any appeals of the established policies and procedures should be directed to the appointing official.

In developing a statewide longitudinal data system, privacy and data protection plans must be in place in each entity that holds student records with personally identifiable information. This includes, for example, preschools, elementary and secondary schools, postsecondary programs and institutions, and workforce training programs. It also includes different organizational levels within each of these components of the education system; for example, elementary and secondary school data are typically held at the school, district, and state levels. Whether they are developed separately at each level or as a part of a unified approach across levels, efforts must be undertaken to ensure that the policies and rules and regulations are compatible across levels. For example, in elementary and secondary education, there may be more information maintained in a student education record at the school and district level than is planned at the state level. In this case, if the privacy and data protection plans are being developed and promulgated from the state level, districts and schools must supplement their plans to ensure that all personally identifiable information maintained about their students is included. On the other hand, if each education level is developing privacy and data protection plans separately, efforts must be undertaken to ensure that established policies and procedures are complementary and do not conflict.

Conduct an Inventory of Personally Identifiable Information

In order to ensure that the necessary data protections are in place, the Governance Committee or a Data Subcommittee for each entity that holds student records must first identify the personally identifiable data elements that need to be protected. Personally identifiable information (PII) includes information that can be used to distinguish or trace an individual's identity either directly or indirectly through linkages with other information. In the case of education data, FERPA regulations (34 CFR § 99.3).

The term personally identifiable information includes, but is not limited to:

1. The student's name;
2. The name of the student's parent or other family members;
3. The address of the student or student's family;
4. A personal identifier, such as the student's Social Security Number, student number, or biometric record;
5. Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;
6. Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty; and/or
7. Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates. (34 CFR § 99.3)

In conducting the inventory, the specific use of PII must be taken into account. For example, while FERPA has provisions to protect students' right to privacy, including the right to inspect and review education records (20 U.S.C. § 1232 (a); 34 CFR § 99.10) and a requirement for consent to disclose information to unauthorized entities (20 U.S.C. § 1232 (b); 34 CFR § 99.30), FERPA permits the release of student directory information[1] (20 U.S.C. § 1232g(a)(5); 34 CFR § 99.3). A school directory may include PII such as a student's name, grade level, and contact information. Taken by itself, the release of this information is not harmful to a student. However, when combined with the student's Social Security Number or another identifier and the student's education record, this information has the potential for violating a student's right to privacy. The release of this combined record could lead to harm or embarrassment. Thus, the privacy and data protection program should focus on PII that will be maintained in the electronic student record system with its likely wealth of student data.[2]

Identify All Personally Identifiable and Sensitive Information

The inventory should include all current and proposed data elements (National Institute of Standards and Technology [NIST], *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* 2010 Special Publication 800-122, pg. 2-2). It should also identify both direct and indirect identifiers. *Direct identifiers* provide information that is unique to the student or the student's family (e.g., name, address, Social Security Number, other unique education-based identification number, photograph, fingerprints, or other biometric record). *Indirect identifiers* are not unique to the student or the student's family but can be used in combination with other information about the student to identify a specific student (e.g., racial or ethnic identity, date of birth, place of birth, mother's maiden name, grade level, participation in a specific program, course enrollment).

An analysis of indirect identifiers should consider the likelihood of identifying an individual student both as a result of a combination of multiple data elements included in the student's education record and as a result of linking the information in education records to information included in external databases. In the first instance, a combination of data elements within student education records might reveal that there is only one student in a specific grade within a school with a set of observable characteristics who experienced a negative academic outcome (e.g., one Hispanic third-grader receiving instruction as an English language learner failed to reach the *proficient* performance level on the state reading assessment). In the second instance, if an external database contains enough overlapping data elements that are unique to an individual student, the two databases can be linked and any additional PII included in the external database can then be associated with that student's education record.

Linkage with information from an external source could occur as a result of a direct linkage by someone with access to two confidential data systems who is able to directly link the two databases (e.g., the student record linked to local public health records on sexually transmitted diseases or local crime records) or as a result of a less direct linkage of information from a student's education record with information available in public records (e.g., the education record for a 15-year-old Asian female includes participation in services for unmarried pregnant students, and public birth records could be used to identify the father of the student's child. Alternatively, an education record might show that a 13-year-old female student was the victim of a violent assault during the school day on a specific date (without the specifics of the assault). Meanwhile, a report in a local newspaper, while protecting the direct identifiers of the victim, reveals some of the details of an assault on a female student in that school on the same date).

At the elementary and secondary level, an analysis of the indirect identifiers should also consider whether any of the data elements are protected under the Protection of Pupil Rights Amendment (PPRA) (20 U.S.C. § 1232h; 34 CFR § Part 98). To protect the privacy and related rights of

[1] Educational agencies or institutions are granted the authority, under FERPA, to publicly release directory information after providing public notice to the parents of students or to eligible students in attendance at the agency or institution of the types of personally identifiable information that the agency or institution has designated as directory information. The parent or the eligible student must also be given the right to refuse to have any or all of the student's information released as directory information.

[2] An electronic student record system, or information system, consists of a discrete set of information resources organized for the collection, processing, maintenance, use, sharing, dissemination, or disposition of [education] information. (44 U.S.C. § 3502)

students and parents, the PPRA requires written parental consent before a minor student can be required to participate in any survey, analysis, or evaluation funded by the U.S. Department of Education that includes information concerning the following:[3]

1. Political affiliations or beliefs of the student or parent;
2. Mental and psychological problems of the student or the student's family;
3. Sex behavior or attitudes;
4. Illegal, anti-social, self-incriminating, and demeaning behavior;
5. Critical appraisals of other individuals with whom respondents have close family relationships;
6. Legally recognized privileged or analogous relationships, such as those of lawyers, physicians, and ministers;
7. Religious practices, affiliations, or beliefs of the student or the student's parent; or
8. Income (other than that required by law to determine eligibility for participation in a program or for receiving financial assistance under such program).

In the event any data elements under consideration for inclusion in a student record system involve any of these eight topics, those data elements should be included on the inventory of PII and should be identified on the list as PPRA-related variables.

A number of data systems include data on students' instructors. A teacher identification number, a student-teacher link, and information on the teacher's education, certification, teaching assignments, and scores on teacher assessments are examples of the types of teacher data elements that may be included at the preschool, elementary, and secondary levels. A faculty identification number, a student-faculty link, and information on the faculty member's field, education, tenure status, credit hours taught in the relevant academic period, and amount of funded research may be included at the postsecondary level. Although FERPA and the definitions given refer specifically to students, PII on teachers and any other staff that are maintained as part of the electronic record system should be included in the inventory of PII and protected in the same way as the student data. Apart from the fact that protecting any PII is a best practice, when faculty and staff data are linked to the student's record, they become indirect identifiers for the student record and can be used to identify individual students.

[3] Under PPRA (20 U.S.C. § 1232h; 34 CFR Part 98), school districts receiving funds from the U.S. Department of Education are required to provide annual parental notification of their policies concerning students' rights and of the specific or approximate dates during the school year of any survey that is scheduled to be administered to students if the survey includes any of the eight restricted topics, regardless of survey funding.

Confirm the Need to Maintain Personally Identifiable Information

The Fair Information Practice of *Data Minimization and Retention* calls for "only collecting personally identifiable information that is directly relevant and necessary to accomplish the specified purpose(s). [And for] only retaining personally identifiable information for as long as is necessary to fulfill the specified purpose(s)." In addition, the Fair Information Practice of *Purpose Specification* calls for "…specifically articulating the purpose or purposes for which the PII is intended to be used." Once the list of current or planned PII in an education record is completed, the planned uses should be identified for each data element (NIST, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* 2010 Special Publication 800-122, pg. 3–4). Decisions should be made as to whether each data element is needed.

The National Forum on Education Statistics[4] identified the following K–12 administrative uses of student education records in the 2004 report *Forum Guide to Protecting the Privacy of Student Information: State and Local Agencies* (pg. 44):

» INSTRUCTION—Teacher and counselors need information about an individual student's previous educational experiences and any special needs the student might have to deliver appropriate instruction and services and to plan educational programs; parent contact information is needed to keep parents informed of student progress.

» OPERATIONS—Schools and districts need data for individual students to ensure the efficiency of day-to-day functions such as attendance records, meeting individual students' special needs, handling individual students' health problems, and operating food service and transportation programs.

» MANAGEMENT—Schools, districts, and state education agencies use data about students for planning and scheduling educational programs and for the distribution of resources.

» ACCOUNTABILITY—Schools, districts, and state education agencies use data about students and about individual students' progress to provide information about students' accomplishments and the effectiveness of schools and specific educational programs.

» RESEARCH AND EVALUATION—Schools, local, state, and federal education agencies use data about students and about individual students' progress to conduct analysis of program effectiveness, the success of student subgroups, and changes in achievement over time to identify effective instructional strategies and to promote school improvement.

Recent legislative initiatives provide funds for states to develop and implement statewide longitudinal data systems to support data-driven decisions to improve student learning and to facilitate research to increase student achievement and close achievement gaps.[5] These data systems are intended to enhance the ability of states to manage, analyze, and use education data. The supporting legislation calls for an expansion in the amount of information included in student education records, including linkable student and teacher identification numbers and student and teacher information on student-level enrollment, demographics, program participation, test records, transcript information, college readiness test scores, successful transition to postsecondary programs, enrollment in postsecondary remedial courses, and entries and exits from various levels of the education system. To facilitate the usefulness of this information, the legislation also calls for an alignment between P–12 and postsecondary data systems, which requires linkages between student and teacher records, between preschool and elementary education, and between secondary and postsecondary education and the workforce.[6] These linkages require data sharing across different components of the education system.

---

[4] This entity is a part of the National Cooperative Education Statistics System, which is authorized in law (20 U.S.C. § 9547). It was established and is supported by the National Center for Education Statistics for the purpose of assisting in producing and maintaining comparable and uniform information and data on early childhood education and elementary and secondary education. To this end, the National Forum proposes principles of good practice to assist state and local education agencies.

[5] Educational Technical Assistance Act of 2002, Title II of ESRA, 20 U.S C. § 9607.

[6] The America COMPETES Act, 20 U.S.C. § 9871 identifies data elements that are important in statewide longitudinal data systems, Title VIII of the American Recovery and Reinvestment Act of 2009 (ARRA, Pub. L. 111-5), authorizes funds to the Institute of Education Sciences to carry out section 208 of the Educational Technical Assistance Act, $250,000,000, which may be used for Statewide data systems that include postsecondary and workforce information, and Title XIV of this Act requires states accepting funds under this Act to establish statewide longitudinal data systems that incorporate the data elements described in the America Competes Act.

Some of the uses of education data require PII from individual students' records; others use aggregated student data for one point in time that are derived from information included in education records; others use aggregate student data that are derived from longitudinal data on individual students; still others use individual student level data linked across levels of the education system. Thus, some uses require access to PII, including the students' names and contact information, and, in some cases, linked longitudinal data; some may require detailed linked longitudinal data included in student records but do not require access to the individual students' names or other direct identifiers; still others may require nothing more than aggregates of data for a single year, again with no need for any information on individual students. Lists of the specific anticipated uses and linkages of the data can help to clarify data needs and to identify those needs which do or do not require access to PII. In addition, given the utility of linking data across sectors, care should be taken to ensure that the direct identifiers that are needed for accurate linking across record systems are maintained.

The length of time student records are retained is complicated by the fact that students may need to request information from education records as proof of credentials for employment purposes over the course of their workforce careers. To protect student privacy, while at the same time maintaining student records, the Governance Committee should develop a schedule and plan for migrating student education records to a retrievable archive following a student's completion at a specific level or departure due to transferring or dropping out. This would preserve the student education records for use in documenting a student's educational credentials (e.g., grade level and/or courses completed and grades or scores earned, honors conferred) and would allow for linkages across sectors and for retrospective evaluations of educational progress. At the same time, archiving historic student education records in a secure environment that is separate from the currently active components of an electronic student record system decreases the likelihood of unauthorized or inadvertent disclosures of records belonging to former students. Similarly, the Governance Committee should establish a plan for record destruction at such point in time when it is anticipated that the records will no longer be needed.

Ensure Data Quality and Integrity

The Fair Information Practice of *Data Quality and Integrity* calls for "ensuring, to the greatest extent possible, that personally identifiable information is accurate, relevant, timely, and complete for the purposes for which it is to be used." The issue of relevance will have already been addressed in the review of the specific uses and need for individual data items. Once a decision is reached to maintain a specific data element in students' education records, there is an obligation to ensure that the information included is up to date and complete and that it accurately reflects the students' educational experiences. Systems should be put in place to ensure the regular periodic updating of student education records with the most current and accurate information available for the intended purpose (e.g., an annual review and updating of student course transcripts). In fact, in recognition of the importance of these elements of student privacy, FERPA (20 U.S.C. § 1232g (a) and the related regulations (34 CFR § 99) acknowledge the right of a parent to inspect and review his or her child's (or, in the case of an eligible student, his or her own) education record for accuracy and to ensure that there are no violations of privacy with the right to request a correction or amendment.

Identify the Risk Level Associated with Different Types of Personally Identifiable Information

Not all personally identifiable data have the same level of sensitivity.[7] Some personally identifiable data elements are more identifiable and/or more sensitive than others and may thus require more electronic security and more controls on access to the data elements. To guide the organization's use of PII and the protections provided for such data, the Governing Committee or the Data Subcommittee should also evaluate the risk of harm associated with each personally identifiable data element. All PII included in a student education record system must be protected, but some may require additional protections (e.g., Social Security Numbers, disciplinary record, medical records).

[7] Sensitivity should be evaluated both in terms of the specific data element and other available personally identifiable data elements. Note that an individual's SSN, medical history, or financial account information is generally considered more sensitive than an individual's phone number or ZIP code.

PII that is unique to a specific individual is more identifiable than certain other personally identifiable data elements that may be shared with others. For example, a student's Social Security Number, fingerprints, or other biometric data are unique to an individual. In contrast, other personally identifiable data elements, such as a ZIP code or date of birth may be shared by multiple students.[8]

In evaluating the sensitivity of individual personally identifiable data elements, the Governing Committee or the Data Subcommittee should take the potential for harm from an unauthorized or inadvertent disclosure into account. In this context, harm refers to any adverse effects that would be experienced by an individual whose PII was the subject of a loss of confidentiality, as well as any adverse effects experienced by the organization that maintains the PII[9] (NIST, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* 2010 Special Publication 800-122, p. 3-1, 2). In the case of a student, harm might include, for example, identity theft, discrimination, or emotional distress. The related harm to the organization responsible for the confidentiality breach could include loss of public confidence and public reputation, administrative burden of investigating the breach and ensuring necessary remedial steps are taken, and financial losses. To start the process of mitigating the disclosure of harmful information, personally identifiable data elements can be categorized by level of sensitivity (i.e., the likelihood of harm from an unauthorized disclosure)—perhaps as high, medium, and low. Note that any data elements identified as a PPRA-related variable should be categorized as a high-risk data element. After the risk level is established, consideration should be given to providing more protection and more restrictions on access for the data elements that are identified as highly sensitive. For example, these data elements might be stored apart from the rest of the student record in a more secure electronic environment, with access limited to "need to know" circumstances for only a subset of those with access to the system.

Summary

At this point the Governing Committee or its Data Subcommittee has inventoried and listed all personally identifiable data elements. The list includes descriptions of the following for each personally identifiable data element:

- Content/definition;
- Type of identifier—direct or indirect;
- PPRA related variable status;
- Specific use(s) and relevance;
- Accuracy;
- Timeliness for the intended use; and
- High, moderate, or low risk of harm from disclosure.

After a thorough review of the list, the Governing Committee should decide whether to retain all existing personally identifiable data elements and whether to go forward with the inclusion of any additional proposed personally identifiable data elements. The inventory of personally identifiable data should be updated each time new data elements are considered for inclusion in the student record data system.

[8] It is important to note, however, groups of the less sensitive identifiers can be combined to identify specific individuals. For example, researcher Latanya Sweeney used public anonymous data from the 1990 census to show that the combination of the five-digit residential ZIP code, gender, and exact date of birth could likely lead to the identification of 87 percent of the population in the United States (in 2005 testimony before the Pennsylvania House Select Committee on Information Security, House Resolution 351, Recommendations to Identify and Combat Privacy Problems in the Commonwealth).

[9] Harm to an individual includes any negative or unwanted effects (i.e., that may be socially, physically, or financially damaging).

Implement Internal Procedural Controls to Protect Personally Identifiable Information

The Fair Information Practice of *Security* calls for "Protecting personally identifiable information (in all media) through appropriate administrative, technical, and physical security safeguards against risks such as loss, unauthorized access or use, destruction, modification, or unintended or inappropriate disclosure." There are a variety of internal controls that can be employed to assist procedurally in the management of personally identifiable data.[10] The first set is a technical solution that involves assigning new unique student identifiers to protect students' PII in longitudinal electronic data systems. The second set focuses on procedures for workforce security to ensure that only authorized staff members are given access to personally identifiable student records. The third set combines aspects of the first two sets of controls in a role-based management approach that is intended to ensure that access to each student's education record is available on a "need-to know" basis. The fourth set involves operating rules for the conditions of use, such as rules concerning permissible uses and prohibiting unauthorized uses, procedures for protecting PII when it is in the possession of authorized users, and procedures for ensuring destruction of copies of records at the end of a period of authorized use. The fifth set of internal controls involves planning for possible data breaches by establishing procedures for reporting known or suspected breaches, analyzing the causes and impact of breaches, and notifying affected individuals.

Unique Student Identifiers and the Use of Linking Codes as Controls for Sensitive Information

In order to monitor the educational progress and experiences of individual students as they progress through the education system, a unique record identifier is needed to link each student's electronic record across grade levels and across schools, institutions, and related educational programs. Once attached to a student record, this identifier becomes part of that student's PII, as it must be unique to the student to be useful.

Each child already has a unique Social Security Number that could also be used to link to information in a student record system with information from education-related activities in other social service programs (e.g., Head Start or family services); thus, this might seem like the logical number to use as the student identifier in an electronic student record system in a K–12 or postsecondary setting. However, the Social Security Number should be treated as a sensitive piece of PII. In addition to being used to track a number of official electronic transactions, it is the single most misused piece of information by criminals perpetrating identity thefts. Using it on a day-to-day basis in an electronic student record system increases the possibility of a harmful disclosure that has ramifications beyond the student's education record. Instead, a separate unique student identifier that is distinct from the student's Social Security Number should be used on a day-to-day basis in an electronic record system.

The unique student identification number can be assigned at the school, district, or state level; however, care must be taken to ensure that within any record system each student has only one assigned identification number and that two students do not share the same identification number. If student records from separate schools within a district form a district-wide student record system, the student identification numbers should be assigned at the district level to ensure that each student in the district has a single unique identification number. Similarly, if all of the school districts in a state form a state-wide student record system, the student identification numbers should be assigned at the district level to ensure that each student in the state has a single unique identification number.

Each student's Social Security Number should be maintained as a data element in student record system because of the important role it plays when linkages are needed to other record systems (e.g., across states or across education levels within a state); however, consideration should be given to storing the student's Social Security Number in a separate secure location. To link the Social Security Number back to the rest of the student's record, a separate linking code must be assigned to each student's record. By attaching a linking code to each student's record, the student's Social Security Number, any other highly sensitive student information, and a copy of the linking code could

[10] There are also a number of electronic controls that can be implemented to assist in the management of personally identifiable data. They will be covered in a Technical Brief on electronic security.

be stored in a separate secure location apart from the student record that is used on a day-to-day basis. The linking code should not be based on a student's Social Security Number or other personal information, should not be used to identify a student's personal information, and should only be used for linking different components of individual student records.

Only a limited number of staff should have knowledge of the method used to generate the linking code. Further, only a limited number of authorized staff should have access to the secured sensitive information and should be permitted to use the linking code to combine two sets of records. Minimizing the number of times a student's Social Security Number and other sensitive data are accessed and limiting access to this information to a small set of authorized persons can help prevent unauthorized and inadvertent disclosures of the Social Security Numbers and other sensitive data.

Each student record system could use its own unique internal linking codes. Then, when record linkages are needed across different record systems (e.g., between states when a student moves or between a secondary school data system and a postsecondary institution's data system), each system can use its linking code to link the student record to the secured Social Security Number. The record(s) with Social Security Numbers attached should be safely transmitted to the administrator of the receiving record system and then stored in a secure environment until the records from the two separate systems are linked by matching the Social Security Number from the two record systems. Once the linked file is created and the data are checked, the Social Security Number should be removed from the combined file, and each student's linking code and Social Security Number is again securely stored.

Workforce Security and Authorization for Access to Personally Identifiable Information

Students and their parents provide the PII requested by the education system, with an expectation that the confidentiality of the information provided will be protected. To ensure that this expectation is fulfilled, administrators have a responsibility to confirm the trustworthiness of employees to whom sensitive student information is entrusted. This can be done through the use of security screenings, training, and binding confidentiality pledges.

PII carries a potential for misuse. As a result, it is advisable to require security screenings for staff members whose job responsibilities require them to have access to PII in student education records. The screening might include a background investigation using written, electronic, telephone, or personal contact to determine the suitability, eligibility, and qualifications of a staff member for employment.[11]

Administrators should establish job descriptions that delineate any uses of information that require access to PII from student education records Administrators should then provide annually recurring training to inform each employee with any job responsibilities that involve student education records of all legal and regulatory safeguard requirements that apply to the use and the design, development, operation, or maintenance of electronic student education records. The training should also cover all rules and procedures that are in place to ensure compliance with the safeguard requirements. Finally, the training should inform employees of the penalties that apply to the misuse of the information in student education records (NIST, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* 2010 Special Publication 800-122, p. 4-1, 2, 3).

Following training, signed Affidavits of Nondisclosure can be used when providing access to confidential data to help ensure awareness of and compliance with all laws, regulations, rules, and procedural protections that apply. The affidavit should include the following:

- The time period approved for access;
- A pledge to protect the personally identifiable data in each student's education record;
- Citations to relevant laws, regulations, and rules;
- A description of penalties for violations; and
- An affirmation that the staff member has read and is aware of the documentation of the rules for handling and using student education records.

[11] The U.S. Department of Education requires all staff and contractors with access to personally identifiable information to undergo a security screening.

Requiring each authorized staff person to sign an Affidavit of Nondisclosure prior to being granted access to student education records fulfills the confidentiality pledge function.

Affidavits of Nondisclosure can be maintained to provide a record of the fact that each authorized staff member affirmed his or her commitment to protect the PII in student education records.

Once the affidavit is in place and access is granted, there are additional electronic mechanisms that can be used to protect the student education records and to monitor and record access and use for auditing and accounting purposes. Electronic security will be addressed in a separate Technical Brief.

Role Based Access to Student Record Data

As mentioned briefly in the discussion of job descriptions, the student information needed on a day-to-day basis varies across groups of employees depending on their roles in the education system. For example, an elementary school teacher is likely to need regular access to student data on attendance, grades, and student performance on various assessments, but not necessarily access to detailed information on the student's medical history or prior disciplinary actions. There are also likely to be differences in the amount of PII needed across levels of the education system. A program administrator for a district-wide program with a specific emphasis, such as science, math, or the arts, would need access to student education records including academic history and students' direct identifiers to organize placements into such programs. Meanwhile, an analyst in the district office who is responsible for generating aggregated reports of student performance for submission to the state education agency would need access to the performance results but not the direct identifiers for individual students.

Once defined, the job descriptions can be used to identify sets of data elements that are needed by groups of data users based on their roles in the education system. Then, rather than allowing each employee access to the full electronic student record or restricting access to needed data elements one user at a time, the database manager grants access to a set of data elements based on the data user's role.

This has been operationalized in statewide student record systems by the use of different access levels to protect personally identifiable and sensitive information in students' records. The Missouri Student Information System documentation, *Data Access and Management Policy* (pg. 6), offers a clear description of the goals in using access levels in the following statement: "All access levels are assigned in a way that maximizes usage by educators without risking inappropriate disclosure of personally identifiable information" http://www.dese.mo.gov/MOSIS/.

When a state uses access levels to control access to information in student records, the access level may control access to full records, with teachers, for example, being limited to students in their assigned classes, and principals having access to all student records in the school. The access level may also be used to control access to specific data elements (or fields) in the student records; finally, access levels can also be used to limit access to read only or to allow read and write access. In some instances, these three dimensions of control are used in combination (e.g., giving a teacher read and write access to a subset of data elements in the student records for the students enrolled in the teacher's class). As states develop systems for sharing student records across levels of the education system, the use of access levels can be expanded to encompass different roles in data use across levels.

Using Education Records

Once staff members have been authorized and granted access to student education records, they must abide by established rules and procedures for using the data—consistent with the terms agreed to in the Affidavit of Nondisclosure. Many of the security controls involved in using the data will be discussed in the Technical Brief on electronic security. However, there is an interface between access and use procedures and electronic security. Specifically, the Governance Committee should establish rules that identify where student records can be accessed. Within the school or office there may be restrictions placed on where staff members can access electronic student records. For example, access to the most sensitive information might be limited to specified secure locations, while access to less sensitive information might be allowed on a wider range of terminals. There may also be restrictions on whether access to student records is limited to the school or office, or whether remote access is permitted.

The use of access restrictions among authorized users will help protect the information in student records from authorized users who might be tempted to look at information they are not authorized to access (i.e., browsing) or from other unauthorized uses of student data. However, even among the staff members granted access to student records use of the information should be limited to permissible uses for the individual data elements, as established in the data inventory. To reinforce this, the Governance Committee should promulgate rules that prohibit browsing or unauthorized uses of information included in student education records.

The Governance Committee should also identify specific behaviors that could lead to inadvertent unauthorized access and establish rules prohibiting these actions. For example, authorized data users should not share a computer that houses identifiable student records with anyone not authorized to access those records, and they should not leave student record data with PII on an unattended computer screen. In a similar vein, if staff members are authorized to print hard copy of PII from student records, there should be rules that require secure storage of hard copy printouts or records (i.e., in a locked cabinet). In addition, if staff members are authorized to copy PII from student records to a CD-ROM or flash drive, there should be rules concerning security and protection of these electronic devices. There should also be record retention rules that govern the length of time a staff member may maintain a local electronic copy or subset of student record data and the length of time that a staff member can maintain hard copy of PII from student records. There should be complementary rules and procedures that govern the destruction of electronic and hard copy extracts of student information at the end of the approved access period.

Breaches of Personally Identifiable Information

Every privacy and data protection plan should include a response plan for the appropriate handling of a breach of PII if one occurs. The NIST 2010 *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* includes a detailed discussion of how to handle data breaches. In particular, the Governance Committee should develop a clear description of what constitutes a breach. That description should be communicated to all staff members who are authorized to access PII in student records, along with a description of the immediate steps to take in the event a security breach occurs or is suspected. In particular, there should be a designated person in the management chain to notify in the event of known or suspected breaches involving PII. Contact information for the designated manager should be disseminated to all staff members, along with a list of the information that should be provided when reporting a known or suspected breach. The NIST 2010 Guide (Special Publication 800-122, pg. 5-1, 2) recommends that the report should include the following information:

- The name, job title, and contact information of the person reporting the incident;
- The name, job title, and contact information of the person who discovered the incident;
- Date and time the incident was discovered;
- Nature of the incident (e.g., system level electronic breach, an electronic breach of one computer or device, or a breach of paper extracts of records);
- Description of the information lost or compromised;

- Name of electronic system and possible interconnectivity with other systems;
- Storage medium from which information was lost or compromised;
- Controls in place to prevent unauthorized use of the lost or compromised information;
- Number of individuals potentially affected; and
- Whether law enforcement was contacted.

Known or suspected breaches of PII from student records should be reported as quickly as possible in an effort to mitigate any adverse events resulting from the breach. The Governance Committee should establish a time span for the reporting requirement (e.g., within one hour of discovery). The Governance Committee should also identify in advance how, when, and to whom notifications should be made (e.g., law enforcement, financial institutions, affected individuals, media, the public). Decisions concerning the breach notification should also be made as to the following:

- Whether breach notification to affected individuals is required;
- Timeliness of the notification;
- General content of the notification;
- Source of the notification (e.g., principal, superintendent, school board);
- Means of providing the notification (e.g., letter or public announcement);
- Who receives the notification (e.g., only affected individuals, general public);
- Remediation options to be provided, if any (e.g., a free copy of credit report, credit monitoring); and
- What corrective actions were taken and by whom.

When a breach occurs, the designated authority should conduct an analysis of the likelihood of exposure and potential harm to affected individuals (e.g., in the case of student records did the breach include Social Security Numbers and other information that could be used in identity theft, or was it limited to PII about the affected students' educational performance). This analysis will inform whether notification is required and the content of breach notification that is provided to affected individuals. There should also be an analysis of the circumstances that resulted in the breach so that the system or procedures can be modified as quickly as possible to avoid further breaches through the same mechanism.

Summary

At this point, the Governing Committee or its Data Subcommittee has reviewed job descriptions and identified the data elements needed for each position, identified authorization procedures for individual staff, and developed rules of access for authorized staff. The Governing Committee or a subcommittee has established a set of procedures to be used to assign unique student identification numbers for day-to-day use and has decided on a specific system architecture to be used in managing access to specific data elements. The Governing Committee or a subcommittee has also promulgated rules specifying the conditions of use for information in student education records, identifying permissible uses and prohibiting unauthorized uses; they have also established procedures for protecting PII when it is in the possession of authorized users and procedures for records disposition. Finally, the Governing Committee has also developed a plan of action to be executed in the event of a data breach.

Provide Public Notice of Education Record Systems

Providing public notice of the existence and use of a student education record system is another essential component of a privacy and data protection program. The Fair Information Practice of *Transparency* calls for "providing notice to the individual regarding the collection, use, dissemination, and maintenance of personally identifiable information" (NIST 2010 Special Publication 800-122, p. D-2, 3).

Annual Notifications

Consistent with the Fair Information Practice of transparency, FERPA and the related regulations require each educational agency or institution that receives funds from the U.S. Department of Education to provide all parents or eligible students[12] an annual notice of their rights with regards to the existence and use of student education records (20 U.S.C. § 1232g(e), 34 CFR 99.7). Insofar as some direct student identifiers are made available publicly as Directory information, FERPA also requires that parents are given an annual notice of the school or districts definition of student directory information, with the opportunity to opt out of the inclusion of their child's, or the eligible student's, directory information (20 U.S.C. § 1232g (e), 34 CFR § 99.7).

FERPA

Under FERPA and the related regulations, the institution, school, or the school district must provide parents with annual notification of their rights[13] and the procedures to use to inspect and review their children's education records and to seek amendment of inaccurate or misleading information in that record.[14] Furthermore, parents must be notified of the disclosures that are permissible under law without their consent,[15] and of the fact that they must consent to other disclosures of PII from their children's education records. Finally, the annual FERPA notice must describe the procedure for a parent to follow in filing a complaint of an alleged violation with the Family Policy Compliance Office (FPCO) in the U.S. Department of Education.

The annual notification does not have to be made individually to parents. Instead, it can be done through any of the following: local or student newspaper, calendar, student programs guide, rules handbook, or other reasonable means.

Directory

A school or district is also required to provide an annual Directory notice, if directory information is disclosed without consent. The school or district may choose to combine their annual FERPA notification with their annual Directory notice. Directory information includes information contained in a student's education record that would not generally be considered harmful or an invasion of privacy if disclosed. The Directory notice must describe the specific types of information the school or district has designated as directory information, and the parent's right to opt out of disclosure of directory information. In the case of postsecondary institutions, these rights accrue to the student.

PPRA

The Pupil Protection Rights Act requires parental notification if a study to be conducted in a school includes any information or questions about the student or the student's family related to the eight identified sensitive topics: political affiliations or beliefs; religious practices, affiliations, or beliefs; mental and psychological problems; sex behavior or attitudes; illegal, anti-social, self-incriminating

[12] Eligible students are those age 18 and over or enrolled in postsecondary institutions.

[13] These rights transfer to the student when he or she turns 18 years of age or enters a postsecondary educational institution at any age ("eligible student").

[14] These requirements are consistent with The Fair Information Practices of Individual Participation and Redress, where redress involves "providing mechanisms for appropriate access, correction, and redress regarding the use of personally identifiable information."

[15] This must include a description of who is considered to be a school official and what is considered to be a legitimate educational interest.

and demeaning behavior; critical appraisals of family members; legally recognized privileged relationships; or income.[16]

If the study is funded by the U.S. Department of Education, schools and contractors must obtain written parental consent before minor students can be required to participate in the study. If the school received funds from the U.S. Department of Education, school districts are required to provide an annual schedule of the specific or approximate dates of all other surveys with a notification of the parents' right to request and review a copy of the survey before it is administered and to decide that their child will not participate, regardless of the survey's source of funding. Under this Act, parents must also be notified each year of their right to decide whether or not their child will participate in activities that make student's personal information available for marketing or other profit-making activities.[17] Parents must also be notified of their right to decide whether or not their child will participate in any non-emergency, invasive physical examination or screening that is scheduled in advance and administered by the school as a required condition of attendance but that is not necessary to protect the immediate health and safety of students.

Under PPRA, schools and contractors are also required to make instructional materials that will be used in any of the studies in which their children participate available for the parents' inspection. Planned surveys that include protected information must be made available for the parents' inspection prior to the administration of the survey.

Resources

The FPCO website includes more specific details and model FERPA notices to use at the school or district level (http://www2.ed.gov/policy/gen/guid/fpco/ferpa/lea-officials.html) and at the postsecondary institution level (http://www2.ed.gov/policy/gen/guid/fpco/ferpa/ps-officials.html), as well as a model Directory notice (http://www2.ed.gov/policy/gen/guid/fpco/ferpa/mndirectoryinfo.html) and a model PPRA notices for use by school districts (http://www2.ed.gov/policy/gen/guid/fpco/ppra/modelnotification.html).

Disclosure of Education Records

The Fair Information Practice of *Individual Participation* calls for "involving the individual in the process of using personally identifiable information and seeking individual consent for the collection, use, dissemination, and maintenance of personally identifiable information." Consistent with this practice, parent's rights to consent to disclosures of PII included in the student's education record must be described in the annual FERPA notice (FERPA, 20 U.S.C. § 1232g (e), 34 CFR §§ 99.7 and 99.30). To meet this requirement, a school must:

- Have a parent's consent prior to the disclosure of education records; and
- Ensure that the consent is signed and dated, specify the records that may be disclosed, state the purpose of the disclosure, and identify to whom the disclosure may be made.

The Fair Information Practice of *Purpose Specification* stresses the importance of "specifically articulating the authority that permits the collection of personally identifiable information and specifically articulating the purpose or purposes for which the personally identifiable information is intended to be used." The annual FERPA notice provides information about permissible uses of PII in education records. That is, FERPA allows educational agencies and institutions to non-consensually release education records to school officials and other designated entities with legitimate educational interests 20 U.S.C. § 1232g(b(1)(A), but the FERPA regulations require educational agencies or institutions that elect to disclose education records to the entities authorized in the Act to use the annual notice to specify the criteria used for identifying a school official and the definition of a legitimate educational interest. Specifically,

[16] See the earlier section *Identify All Personally Identifiable and Sensitive Information* for the complete text of the list as specified in law.

[17] This does not apply to information collected from students to support educational products or student services such as postsecondary education or military recruitment; book clubs, magazines, and programs providing access to low-cost literacy products; curriculum and instructional materials; tests and assessments used to provide information about students; the sale by students of products or services to raise funds for school-related or education-related activities; and student recognition programs.

under the FERPA regulations at 34 CFR § 99.31, a school may disclose PII from education records without consent when:

» The disclosure is to school officials who have been determined to have legitimate educational interests;

  - The disclosure is to other school officials, including teachers, within the agency or institution who have legitimate educational interests; a third-party contractor, consultant, volunteer, or other party to whom an agency or institution has outsourced institutional services for which the agency or institution would otherwise use employees—as long as that third party's use and maintenance of education records is under the direct control of the agency or institution and is subject to the regulation requirements governing the use and redisclosure of PII from education records (34 CFR § 99.33(a)); and
  - An educational agency or institution uses reasonable methods to ensure that school officials obtain access to only those education records in which they have legitimate educational interests (34 CFR § 99.31(a)(1));

» The disclosure is to officials of another school, district, or institution of postsecondary education where the student seeks or intends to enroll, or where the student is already enrolled so long as the disclosure is for purposes related to the student's enrollment or transfer (34 CFR §§ 99.31(a)(2) and 99.34);

» The disclosure is to authorized representatives of the Comptroller General of the United States, the Attorney General of the United States, the Secretary of the Department of Education, or state and local educational authorities for the purpose of auditing or evaluating federal or state supported education programs or enforcing federal laws which relate to those programs (34 CFR §§ 99.31(a)(3) and 99.35);

» The disclosure is in connection with financial aid for which the student has applied or which the student has received if the information is necessary for such purposes as to determine eligibility, the amount, the conditions for the student to apply for or receive financial aid or enforce the terms and conditions of the aid (34 CFR § 99.31(a)(4));

» The disclosure is to organizations conducting studies for, or on behalf of, educational agencies or institutions for specified purposes related to predictive tests, student aid programs, or the improvement of instruction(34 CFR § 99.31(a)(6));

» The disclosure is to accrediting organizations to evaluate accreditation status (34 CFR § 99.31(a)(7));

» The disclosure is pursuant to a court order or a lawfully issued subpoena[18] (34 CFR § 99.31(a)(9));

» The disclosure is in connection with a health or safety emergency (34 CFR §§ 99.31(a)(10) and 99.36);

» The information disclosed has been appropriately designated as directory information by the school (34 CFR § 99.31(a)(11) and 99.37); and

» The disclosure is of de-identified student level data for the purposes of education research (34 CFR § 99.31(b)).

The SLDS Technical Brief on data sharing agreements will cover recommended terms for inclusion in agreements, along with a discussion of the specific uses permitted under legitimate educational interests, education research, and uses related to predictive tests, student aid programs, and the improvement of education.

Summary

A privacy and data protection program for student education records must include an array of rules and procedures for protecting PII held in the record system. It also must include a full set of public disclosures of the existence and uses of the information included in the data system, a description of all parents' or eligible students' rights to review and appeal the contents of an individual education record and of their rights and the procedures to appeal a violation.

[18] See 34 CFR § 99.31 for additional disclosures related to legal matters.

Accountability and Auditing

The Fair Information Practice of *Accounting and Auditing* calls for "Auditing for the actual use of personally identifiable information to demonstrate compliance with established privacy controls." This involves auditing the use of PII to demonstrate compliance with an organization's privacy and data protection plan, the privacy principles embodied in the Fair Information Practices, and all applicable privacy protection laws, regulations, and administrative requirements. The specific activities to be audited should be identified in the privacy and data protection plan. Many elements of a data security audit involve electronic security and will be discussed in the Brief on that topic. However, there are a several aspects of data stewardship that should be audited to confirm that required actions are taken to ensure the proper use and protection of PII in student education records. A failure to comply with any of the identified auditable elements of the privacy and data protection plan should be reported to appropriate officials for action.

Audit the Inventory of Personally Identifiable Information

The inventory of PII should include all current and proposed data elements (NIST, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* 2010 Special Publication 800-122, pg. 2-2). The data manager should maintain records of the inventory of PII.

In the first data stewardship privacy audit, the inventory should be examined against the content of the existing longitudinal data system to determine whether the list of personally identifiable data elements maintained for students, teachers, and other staff members is complete.

Next, the audit should confirm that the inventory includes all of the required information for each data element. That is, for each data element, the inventory should include an indication of specific uses, whether it is a direct or an indirect identifier and the associated risk level and whether it involves any of the restricted topics identified in the Protection of Pupil Rights Act. Subsequent audits should identify updates to the record system that added new data elements and ensure that each new data element was added to the inventory and that all of the required information is included for each data element.

Audit of Data Quality and Integrity

FERPA (20 U.S.C. § 1232g (a) and the related regulations (34 CFR § 99) establish the right of a parent to inspect and review his or her child's (or in the case of an eligible student his or her own) education record for accuracy. The data manager should develop procedures that result in data that are up to date and complete and that accurately reflect the students' educational experiences. Periodic audits of data quality can support data quality by either substantiating the quality of individual data elements or identifying inaccuracies for correction. Periodic quality audits should be built into the data collection, reporting, and release cycle.

The NCES-sponsored National Forum on Education Statistics published the 2004 report *Forum Guide to Building a Culture of Data Quality* to assist schools and school districts in the development of procedures to improve the accuracy, utility, timeliness, and security of data in education data systems. The Forum web site also provides lesson plans as part of the *Forum Curriculum for Improving Education Data (*http://nces.ed.gov/pubs2007/curriculum/index.asp). The curriculum is designed for use in schools and school districts to support the production of "high-quality education data," with the goal of presenting the concepts and skills needed to improve data quality. One of the lessons included in the curriculum is *Validating and Auditing Data* (http://nces.ed.gov/pubs2007/curriculum/ls_validating.asp).

The goals of the curriculum on data validation and audits include describing the steps required to validate data, describing the purpose of a data audit, and identifying the steps included in a data audit in order to outline a plan for a data audit. The data validation involves data entry, checking for errors, confirming errors are real and not outliers, identifying each place the incorrect data element is stored in the data system, and providing corrections to the data entry staff.[19]

[19] While these data validation activities have broader utility than those involved with privacy, ensuring the accuracy and validity of data maintained in an education record system is consistent with the FERPA requirement that parents have the right to review the accuracy of their children's information.

The audit confirms the accuracy of the data that are released for use by the school and district staff and by the public. To conduct a successful audit of data accuracy, the first step is to identify the released data (e.g., printed reports, tables published on the web, online table generator), and then the data should be analyzed, looking especially for data anomalies. If suspected data anomalies are identified, the audit next focuses on whether they represent real change or whether they are the result of an error. If an error is identified, the source of the error should be investigated (e.g., data recording error, transposed number, data entry error), and the needed correction should be identified. Related procedures are reviewed to identify any needed changes. Staff who contributed to the error should be notified and provided instruction needed to avoid repeating the error. Finally, notice of the changed data should be provided to all data users.

Audits of Internal Controls used to Protect Personally Identifiable Information

Unique Student Identifiers

Longitudinal student record data requires a unique record identifier for each student in a data system. That unique identifier is needed to link each student's electronic record across grade levels and across schools, institutions, and related educational programs. Once attached to a student record, this identifier becomes part of that student's PII, as it must be unique to the student to be useful. Thus, the audit of internal controls should start with an examination of the process used to assign unique student identification numbers. The first question is whether unique identification numbers other than the students' Social Security Numbers are in place for use in day-to-day operations. If so, the next task is to confirm that the student identification numbers are not based on the students' Social Security Numbers; that the students' Social Security Numbers are securely stored apart from the student records that are used daily; that a linking code exists to be used to link a student's record to that student's Social Security Number when the need arises (e.g., the student transfers out of state or transitions to postsecondary education); and that the method for generating the linking key is closely protected, with knowledge limited to a small number of staff positions.

The student identification numbers should be audited to ensure that each student has only one identification number. This can be done electronically by searching for matching data on the combination of name, age, grade, sex, and race/ethnicity. If matches occur, the student records should be examined further to confirm that there are not multiple records for an individual student. These matches should include options for multiple spellings of names and for the use of initials in addition to, or in place of, the first name. If any students are found with multiple student identification numbers, the records should be consolidated into one record using only one of the identification numbers for that student and the duplicate records should be deleted.

Conversely, the student identification numbers should be examined to confirm that the same number is not being used for multiple students. This can be done by electronically searching for exact matches on two or more identification numbers. If matches occur, the associated the records should be examined to confirm whether the records are for different students or whether there are two records for the same student (perhaps with a full first name on one record and initials in place of the first name of the second record). If one identification number has been assigned to two or more students, each student should be given a new unique identification number. If one identification number is being used for two different records for the same student, the two records should be reconciled and combined under the existing student identification number.

Workforce Security and Permitted Access to Personally Identifiable Information

To ensure that the requirements of FERPA are met and that PII is protected, administrators have a responsibility to protect access to that information and to confirm the trustworthiness of employees to whom sensitive student information is entrusted. An audit of workforce security should start with a review of job descriptions to ensure that the need for access to PII is clearly specified. Then once the positions with a need for access are identified, the audit should review the list of staff members in those positions against the documentation for completed background investigations to ensure

that each staff member with access to personally identifiable and sensitive student information has successfully passed a background check. The audit should review the same list of staff members against the list of staff who completed the required privacy and data protection training and the file of signed confidentiality pledges (i.e., affidavits of nondisclosure) to ensure that each staff member with access to personally identifiable and sensitive student information is aware of the relevant laws, regulations, and rules and has agreed to uphold them to protect student information.

The data manager should also have records documenting the authorized level of access for each data user granted access to any personally identifiable student information. There should be an access control system in place, and an audit should be conducted to ensure that each data user's level of access is in line with that person's current job description. If discrepancies are found, the level of access should be corrected, or a justification for the deviation from established access levels should be documented. In addition, the current levels of access should be compared to the approved levels of access. If discrepancies are found, the level of access should be corrected, or a justification should be provided and the data user's access level should be corrected in the data manager's records.

Summary

A privacy and data protection program for student education records must include a set of checks and balances to ensure that the necessary rules and procedures are in place and that they are being fully implemented. This is best done through a formal periodic audit of the various processes involved in the processing and usage of personally identifiable student information. Starting with the careful identification of the personally identifiable and sensitive data elements, continuing through the data processing and reporting to the day-to-day usage of student information. The audit starts by identifying the relevant governing rules and procedures, examines the records for deviations from the rules and procedures, and ensures that needed corrections are implemented. Where possible, the audit should identify the factors that contributed to the problems identified, examine the related processes, and make suggestions for procedural changes that might reduce the number of similar problems in future audits.

References

American Statistical Association, Committee on Privacy and Confidentiality, *Key Terms/ Definitions in Privacy and Confidentiality.* Alexandria, VA: Retrieved from http://www.amstat.org/committees/pc/keyterms.html on 6/17/2010.

Code of Federal Regulations, Title 34—Education, Part 99. *Family Educational and Privacy Rights,* (34CFR99). Washington, DC: GPO Access e-CFR. Retrieved from http://ecfr.gpoaccess.gov/cgi/t/text/ext-idx?c=ecfr&sid=44d350c26fb9cba4a156bf805f297c9e&tpl=/ecfrbrowse/Title34/34cfr99_main_02.tpl on 9/10/2010.

The Federal Chief Information Officers Council (2008). *Federal Enterprise Architecture Security and Privacy Profile, Version 2.* Washington, DC: Federal Enterprise Architecture Program Management Office, Retrieved from http://www.cio.gov/Documents/Security_and_Privacy_Profile_v2.pdf on 6/17/2010.

National Forum on Education Statistic (2004). *Forum Guide to Protecting the Privacy of Student Information: State and Local Education Agencies,* (NCES 2004-330). Washington, DC: Retrieved from http://nces.ed.gov/pubs2004/2004330.pdf on 6/17/2010.

National Forum on Education Statistic (2004). *Forum Guide to Building a Culture of Quality Data: A School & District Resource,* (NFES 2005-801). Washington, DC: Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005801 on 6/17/2010.

McCallister, E., Grance, T., and Scarfone, K. (2010). *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII): Recommendations of the National Institute of Standards and Technology* (NIST Special Publication 800-122). National Institute of Standards and Technology, U.S. Department of Commerce. Washington, DC: Retrieved from http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf on 5/4/2010.

Sweeney, Latanya. (2005). *Recommendations to Identify and Combat Privacy Problems in the Commonwealth.* Testimony on House Resolution 351, Pennsylvania House Select Committee on Information Security.

U.S. Code, Title 20—Education, Chapter 31—General Provisions Concerning Education, Subchapter III—General Requirements and Conditions Concerning Operation and Administration of Education Programs: General Authority of Secretary, Part 4—Records, Privacy, Limitation on Withholding Federal funds, Section 1232g. *Family Educational and Privacy Rights,* (20USC1232g). Washington, DC: GPO Access. Retrieved from http://frwebgate4.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=799486197532+0+1+0&WAISaction=retrieve on 9/10/2010.

U.S. Code, Title 20—Education, Chapter 70—Strengthening and Improvement of Elementary and Secondary Schools, Subchapter I—Improving the Academic Achievement of the Disadvantaged, Part A—Improving Basic Programs Operated by Local Educational Agencies, Subpart 1—Basic Program Requirements, Section 6311. *State Plans,* (20USC6311). Washington, DC: GPO Access. Retrieved from http://frwebgate2.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=bULwJH/21/1/0&WAISaction=retrieve on 9/10/2010.

U.S. Code, Title 20—Education, Chapter 76—Education Research, Statistics, Evaluation, Information, and Dissemination, Subchapter I—Education Sciences Reform, Section 9547. *Cooperative Education Statistics Systems,* (20USC9547). Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/usc.cgi?ACTION=RETRIEVE&FILE=$$xa$$busc20.wais&start=10271732&SIZE=977&TYPE=TEXT on 9/10/2010.

U.S. Code, Title 20—Education, Chapter 76—Education Research, Statistics, Evaluation, Information, and Dissemination, Subchapter II—Educational Technical Assistance, Section 9607. *Grant Program for Statewide, Longitudinal Data Systems,* (20USC9607). Washington, DC: GPO Access. Retrieved from http://frwebgate3.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=FKr6BA/0/1/0&WAISaction=retrieve on 9/10/2010.

U.S. Department of Commerce, National Institute of Standards and Technology (2010). *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* (SP 800-122). Gaithersburg, MD.

U.S. Department of Health and Human Services, Report of the HEW Secretary's Advisory Committee on Automated Personal Data Systems (1973). *Records, Computers and the Rights of Citizens,* Washington, DC: Retrieved from http://aspe.hhs.gov/datacncl/1973privacy/tocprefacemembers.htm on 5/11/2010.

U.S. Department of Homeland Security, Privacy Policy Guidance Memorandum (2008). *The Fair Information Practice Principles: Framework for Privacy Policy at the Department of Homeland Security,* Washington, DC: Retrieved from http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf on 6/17/2010.

U.S. Public Law, 110-69, America Competes Act, Title VI—Education, Section 6401. Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ069.110 on 9/10/2010.

U.S. Public Law, 111-5, American Recovery and Reinvestment Act, Title VIII—Education, Institute of Education Sciences. Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_public_laws&docid=f:publ005.111 on 9/10/2010.

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

SLDS Technical Brief

Guidance for Statewide Longitudinal Data Systems (SLDS)

December 2010, Brief 3
NCES 2011-603

Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting

Contents

SLDS Technical Briefs are intended to provide "best practices" for consideration by states developing Statewide Longitudinal Data Systems.

Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

For more information, contact:
Marilyn Seastrom
National Center for Education Statistics
(202) 502-7303
Marilyn.Seastrom@ed.gov

Introduction

Over the last decade, increased attention on education has led to an expansion in the amount of information on students and their schools and school districts reported to parents and the general public (20 U.S.C. § 6311). States now report student outcomes based on assessments of student achievement in specific subjects and grade levels for all students, as well as for subgroups defined by gender, race and ethnicity, English proficiency status, migrant status, disability status, and economic status. Typically, the data are reported as the percentage distribution of students in a subgroup across achievement levels. These reports are issued at the state, district, and school levels. Additional outcome measures, such as data on attendance, dropout rates, and graduation rates, are also reported frequently.

These reports offer the challenge of meeting the reporting requirements while also meeting legal requirements to protect each student's personally identifiable information (Family Educational Rights and Privacy Act [FERPA]) (20 U.S.C. § 1232g; 34 CFR Part 99). Recognizing this, the reporting requirements state that subgroup disaggregations of the data may not be published if the results would yield personally identifiable information about an individual student (or if the number of students in a category is insufficient to yield statistically reliable information). States are required to define a minimum number of students in a reporting group or subgroup required to publish results consistent with the protection of personally identifiable information (34 CFR § 200.7).

Individual states have adopted minimum group size reporting rules, with the minimum number of students ranging from 5 to 30 and a modal category of 10 (used by 39 states in the most recent results available on state websites in late winter of 2010). Each state has adopted additional practices to protect personally identifiable information about its students in reported results. These practices include various forms of suppression, top and bottom coding of values at the ends of a distribution, and limiting the amount of detail reported for the underlying counts. This Technical Brief includes a summary of key definitions, a brief discussion of background information, and a review and analysis of current practices to illustrate that some practices work better than others in protecting personally identifiable information reported from student education records.

The review led to the formulation of recommended reporting rules that are driven by the size of the reporting groups or subgroups. The reporting rules are intended to maximize the amount of detail that can be safely reported without allowing disclosures from student outcome measure categories that are based on small numbers of students. NCES welcomes input on these recommendations.

Definitions

Personally identifiable information includes the name and address of the student and the student's family; a personal identifier, such as the student's Social Security Number, student number, or biometric record; other indirect information, such as the student's date and place of birth and mother's maiden name; other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of relevant circumstances, to identify a student with reasonable certainty; and information based on a targeted request.

Disclosure means to permit access to or the release, transfer, or other communication of personally identifiable information contained in education records by any means. To avoid disclosures in published tables, whenever possible, data about individual students should be combined with data from a sufficient number of other students to disguise the attributes of a single student. When this is not possible, data about small numbers of students should not be published.

Suppression refers to withholding information from publication. Some information is withheld from publication in a table to protect data based on small counts because the release of the information would likely lead to a disclosure. Other information is withheld from publication in a table to prevent the calculation of the data based on small counts from the published information; this is known as complementary suppression.

Recoding refers to reporting values as being within a specified range rather than as a specific value.

Top coding refers to reporting values over a set value as greater than that value.

Bottom coding refers to reporting values under a set value as less than that value.

Top coding and bottom coding are specific types of recoding. These procedures are used to protect data for individual students from disclosure.

Subgroups refer to students within a larger group who share specific characteristics, such as the subgroup of male students and the subgroup of female students within a school or within a grade in a school. Information from student records is often reported for subgroups of students by gender, race and ethnicity, English proficiency status, migrant status, disability status, and economic status.

Outcome measures refer to the student's educational experiences that are recorded in student's educational records. For example, student grades, courses completed, scores on standardized assessments, school attendance, graduation status, participation in extracurricular activities, and disciplinary actions are commonly reported measures of student outcomes.

Categories refer to groups of students that share specific experiences that comprise the range of possible outcomes for each educational measure. For example, the percent of students with passing as compared to failing grades, the percent of students who dropout as compared to completing high school, or the percent of students who scored at each of several achievement levels on a standardized state assessment.

Background

As the nation has focused its attention on education over the last decade, there has been a large increase in the amount of data reported to the general public on America's students and their schools and school districts (20 U.S.C. § 6311(h); 20 U.S.C. § 9607; U.S. Public Law 110-69; U.S. Public Law 111-5). Reporting requirements for public elementary and secondary institutions that receive federal funds include annual status and progress reports at the school, district, and state levels (20 U.S.C. § 6311(h)).[1] Among other requirements, these reports, identified as report cards, must include results from state assessments on the percent of students assessed, along with student achievement results across achievement levels in specific subjects and grade levels for all students and for reporting subgroups including gender, race/ethnicity, English proficiency status, migrant status, disability status, and economic status. The annual status and progress report cards also typically include data on attendance rates and report graduation rates for secondary schools. Dropout rates are also frequently reported at the district and school levels.

The current reporting requirements are typically met through state-, district-, and school-level reports that are published by each state's department of education. These reports offer the challenge of balancing the reporting requirements against legal requirements to protect each student's personally identifiable information (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99). To this end, the reporting requirements for Title I state that disaggregating the data for specific subgroups may not occur if the number of students in a reporting group or subgroup is insufficient to yield statistically reliable information or if the results would yield personally identifiable information about an individual student (20 U.S.C. § 6311(h); 34 CFR § 200.7).[2]

As part of the reporting requirements, each state is required to have an accountability plan that describes its system for monitoring adequate yearly progress with annual objectives for continuous and substantial improvement for all students and for each specified student subgroup. In addition to defining specific measures, each state's accountability plan is expected to include the state's definition of the minimum number of students in a subgroup required for reporting purposes and information as to how the State Accountability System protects the privacy of students when reporting results.

What does protecting student privacy mean in a reporting context? In order to protect a student's privacy, the student's personally identifiable information must be protected from public release. The broad, federal government-wide definition of personally identifiable information states "the term 'personally identifiable information' refers to information that can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc., alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc." (OMB Memorandum 07-16, *Safeguarding Against and Responding to the Breach of Personally Identifiable Information; Implementation Guidance for Title V of the E Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)*).

[1] The requirement specified in law is for an annual state report card and for annual district report cards that include information for the district and each school.

[2] The law states that reporting student assessment results disaggregated by economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency is not required if the number of students in a *category* is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student (20 U.S.C. § 6311). However, the regulations use the term *subgroup* to refer to the disaggregated student data, and the regulations specify that a state may not report achievement results for a *subgroup* if the results would reveal personally identifiable information about an individual student (34 CFR § 200.7). This is further promulgated in the September 12, 2003 non-regulatory guidance on Report Cards Title I, Part A.

The FERPA definition of personally identifiable information (34 CFR § 99.3) follows the government-wide definition and includes the following:

Personally identifiable information includes, but is not limited to:

1. The student's name;
2. The name of the student's parent or other family members;
3. The address of the student or student's family;
4. A personal identifier, such as the student's Social Security Number, student number, or biometric record;[3]
5. Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;
6. Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty;
7. Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates. (34 CFR § 99.3)

Protecting student privacy means publishing data only in a manner that does not reveal individual students' personally identifiable information, either directly or in combination with other available information. Another way of putting this is that the goal is to publish summary results that do not allow someone to learn information about a specific student.

States publish annual status and progress reports that are based on reports of outcome measures at the school, district, or state level. These reports aggregate, or combine, the results for individual students into summary statistics. These statistics include the number or percentage of students overall or in each of the reporting subgroups for specific outcome measures (e.g., the percentage of students in each racial and ethnic group who graduate from high school; the percentage of English language learners who score in each achievement level on a state assessment).

This report demonstrates how disclosures occur even in summary statistics. It describes various reporting practices and data protection techniques currently in use and illustrates how commonly used methods of data protection may fall short of their goal. The report then identifies "best practices" to avoid the unintended disclosure of personally identifiable information, including publishing the percentage distribution across categories of outcome measures with no underlying counts or totals; publishing a collapsed percentage distribution across categories of outcome measures with no underlying counts or totals; publishing counts but using complementary suppression at the subgroup level when a small subgroup is suppressed; limiting the amount of detail published for school background information; recoding the ends of percentage distributions; and recoding high and low rates. This information is used to develop recommendations for reporting rules that maximize the amount of information reported while protecting the privacy of each student's data.

Unintended Disclosure of Personally Identifiable Information

When personally identifiable information is revealed through information released to the public, it is called a disclosure.[4] When schools, districts, or states release information about educational progress, they typically release aggregated data—data for groups of

[3] FERPA 2008 regulations state that the term "biometric record, as used in the definition of personally identifiable information, means a record of one or more measurable biologic or behavioral characteristics that can be used for automated recognition of an individual. Examples include fingerprints; retina and iris patterns; voiceprints; DNA sequence; facial characteristics; and handwriting." (34 CFR § 99.3)

students—to prevent disclosure of information about an individual. Even with some methods of aggregation, unintended disclosure of personally identifiable information may occur. How could data reporting outcome measures for groups of students possibly reveal information on an individual student? The example that follows shows how information about individual students' achievement levels can be revealed, even in data reported for groups of students. Furthermore, it shows that the identity of groups of students can be revealed within combinations of achievement levels (e.g., Below Basic and Basic for students who scored below Proficient, or proficient and advanced for students who scored at or above Proficient).

Typically, each child's parents are given their child's score and achievement level on the state assessment as well as the report for their child's school. Table 1 provides the percentage distribution and number of students at each achievement level at the school level in grade 4 mathematics, for students overall and for several subgroups: White and Hispanic students, students with and without an individualized education plan, and students who are and are not English language learners. Any combination of these three subgroup variables that reveals the achievement level for a student or group of students with identifiable characteristics results in a disclosure.

Example 1: Unintended Disclosures

Consider a school report that includes results on the state assessment by grade and subject. No results are suppressed as a result of a small subgroup count, since each subgroup included more than the minimum reporting group size of 5. The report shows that there are 32 fourth-graders in this school and that they were all assessed in mathematics (table 1). Among these students, 12.5 percent, or 4 students, scored at the Below Basic achievement level; 31.3 percent, or 10 students, scored at the Basic level; 34.4 percent, or 11 students, scored at the Proficient level; and 21.9 percent, or 7 students, scored at the Advanced level. The data reported for the subgroups of students with and without an individualized education plan show that all fourth-graders with an individualized education plan scored below the Proficient level (4 students at the Below Basic level plus 3 at the Basic level). Assuming that other students in the class know who among their peers have individualized education plans, this is a disclosure because it reveals that each fourth-grader with an individualized education plan failed to reach the Proficient level on the assessment.

Next, looking at the 10 Hispanic fourth-graders, the data show that 1 student in this subgroup scored at the Proficient level, while the other 9 students scored at either the Basic level (5 students) or the Below Basic level (4 students). Since parents receive their child's score and achievement level as well as a school report that shows the performance in mathematics by grade, the parents of the 1 Hispanic student who scored at the Proficient level know that the other 9 Hispanic students in the fourth grade each scored below the Proficient level in mathematics. This is a disclosure, because these parents now know that each of their child's ethnic peers failed to reach the Proficient level.[5]

The subgroup data in this table also show that each of the 4 fourth-graders who scored at the Below Basic level were Hispanic, received English language instruction, and had an individualized education plan. This is a considerable amount of information about the characteristics of the 4 lowest performers. However, since there were Hispanic students who scored at the Below Basic, Basic, and Proficient achievement levels, students with individualized education plans who scored at both the Below Basic and Basic achievement levels, and students receiving English language instruction who scored at both the Below Basic and Basic achievement levels, the table only identifies the fact that there are four Hispanic fourth-graders with this set of three shared characteristics; it does not identify the 4 specific Hispanic students. Thus, the table considered alone does not result in a disclosure in this instance.

Suppose, however, that the students with individualized education plans receive observable special services (e.g., a tutor, extra time on tests, one-on-one test instruction) and that there are exactly 4 Hispanic students receiving these services; then it becomes apparent that these are the 4 Hispanic students who scored at the Below Basic achievement level.

[4] Under FERPA, disclosure means to permit access to or the release, transfer, or other communication of personally identifiable information contained in education records by any means, including oral, written, or electronic means, to any party except the party identified as the party that provided or created the record (34 CFR § 99.3).

[5] While this disclosure is based on the parents' personal knowledge of their child's score, the fact that each parent in the school receives his or her child's score raises this source of disclosure as a topic of concern (i.e., knowledge of one child's score revealing the performance of other students).

Table 1. School-level grade 4 mathematics assessment results in a state with a minimum reporting group size of 5

		Percent assessed	Tested	Below Basic	Basic	Proficient	Advanced
Total	%	100	100	12.5	31.3	34.4	21.9
	N	†	32	4	10	11	7
White	%	100	100	0.0	22.7	45.5	31.8
	N	†	22	0	5	10	7
Hispanic	%	100	100	40.0	50.0	10.0	0.0
	N	†	10	4	5	1	0
Individualized education plan	%	100	100	57.1	42.9	0.0	0.0
	N	†	7	4	3	0	0
No individualized education plan	%	100	100	0.0	28.0	44.0	28.0
	N	†	25	0	7	11	7
English language learner	%	100	100	40.0	50.0	10.0	0.0
	N	†	10	4	5	1	0
Not English language learner	%	100	100	0.0	22.7	45.5	31.8
	N	†	22	0	5	10	7

† Not applicable.
NOTE: Details may not sum to totals because of rounding.

Recall that the reporting requirements acknowledge the risk associated with small numbers by indicating that results should only be published if the results would not reveal personally identifiable information about an individual student. The instructions for the state accountability plan also acknowledge this risk with the requirement for each state to establish a minimum subgroup size for reporting and with the requirement for each state to describe how the State Accountability System protects the privacy of students when reporting results.

Current Disclosure Prevention Practices that Retain Some Disclosure Risk

Typically, a state establishes the required minimum number of students in a subgroup for privacy protection and then does not report the results for outcome measures for any subgroup with less than this established minimum number. The groups not reported are identified as having been suppressed to protect student privacy. A review in late winter of 2010 of the most recent reported assessment results for each state and the District of Columbia found that 39 states use a minimum reporting group size of 10 students. Another 7 states set the minimum reporting group size at 5, and 5 states set the minimum higher, with values ranging from 15 to 30.

While subgroup suppression is a good start, it may not be enough to prevent disclosure of personally identifiable information. The descriptions of current practices include such potentially problematic methods as 1) suppressing data for small subgroups but not for small categories of outcome measures for reported subgroups; 2) suppressing data for small subgroups but reporting counts across the categories of the outcome measure for the overall group and the reported subgroups; 3) suppressing data for small subgroups but reporting the overall total count; and 4) suppressing data for small subgroups but reporting ranges for the overall totals and the reported subgroup totals.

Suppressing Data for Subgroups but not for Reporting Categories

The practice of suppressing data for small subgroups is a start. However, when subgroup results are reported for the categories of an outcome measure, there can also be a small number of students in one or more of the categories within the larger subgroups. Reporting results for small numbers of students within a category or within a subgroup can present a risk to student privacy because it increases the risk of unintentionally releasing information that identifies individual students. The minimum for categories within subgroups can be set lower than the size of the subgroup minimum, but there should be a minimum size specified for individual categories to guard against unintentional disclosures. This minimum, which is sometimes referred to as the threshold rule, defines those categories in a table that are defined as sensitive because the number of students is less than the specified number. Some data collection agencies set this number at 5, while others set it as 3. (Federal Committee of Statistical Methodology, Working Paper 22). Sensitive categories are illustrated in the following example.

Example 2: Suppression of Small Subgroups but not Small Categories

In this example, when a minimum reporting size of 10 is applied to the data from table 1, the assessment results for the 7 students with individualized education plans are presumed to be protected from disclosure because the results are suppressed (see table 2). Thus, the result in example 1 showing that all students with an individualized education plan failed to reach the Proficient level of the state assessment is presumed to be protected from disclosure. However, when the assessment results of the 10 Hispanic students and the 10 English language learners are reported across the four achievement levels, the number of students at each achievement level falls below the established minimum reporting size. In both subgroups, there are 4 students in the Below Basic achievement group, 5 students in the Basic achievement group, and 1 student in the Proficient achievement group; nevertheless, the results are reported since the minimum size rule is applied at the subgroup reporting level. As described in example 1, reporting that only one Hispanic child scored at or above the Proficient level discloses information about that child and about the achievement level of the other students in the subgroup. Anyone who is able to identify the Hispanic child with a high score then knows that the other Hispanic children in the same grade failed to reach the proficient achievement level.

Table 2. School-level grade 4 mathematics assessment results in a state with a minimum reporting group size of 10

		Percent assessed	Tested	Below Basic	Basic	Proficient	Advanced
Total	%	100	100	12.5	31.3	34.4	21.9
	N	†	32	4	10	11	7
White	%	100	100	0.0	22.7	45.5	31.8
	N	†	22	0	5	10	7
Hispanic	%	100	100	40.0	50.0	10.0	0.0
	N	†	10	4	5	1	0
Individualized education plan	%	100	100	*	*	*	*
	N	†	7	*	*	*	*
No individualized education plan	%	100	100	0.0	28.0	44.0	28.0
	N	†	25	0	7	11	7
English language learner	%	100	100	40.0	50.0	10.0	0.0
	N	†	10	4	5	1	0
Not English language learner	%	100	100	0.0	22.7	45.5	31.8
	N	†	22	0	5	10	7

† Not applicable.
* Not reported to protect subgroups with fewer than 10 students.
NOTE: Details may not sum to totals because of rounding.

Suppressing Data for Subgroups but Reporting Too Much Detail in Underlying Counts

Suppressing data for small subgroups is a first step. However, when data are suppressed to protect student privacy, care must also be taken to avoid publishing information that can be used to retrieve or recover the suppressed information. The next three examples illustrate disclosure problems that can occur in reporting student outcome measures.

The released data in each example table are displayed with a white background. The shaded portions of the example tables represent data that were suppressed. The data entries in the shaded portions of the table were recovered from the released data.

Counts for overall group and reported subgroups

In 38 states, the data are suppressed for subgroups that fall below the minimum reporting group size; however, the number of students and the percentage distributions across the categories of the outcome measure are reported for the overall group and the remaining reporting subgroups. The reported information can then be used to recover the suppressed data through a series of calculations. This can be done using the following steps:

1. Convert the percentages across the outcome categories for the overall group to proportions.

2. Multiply the proportions by the number of students in the overall group to yield the number of students in each category of the outcome measure in the overall group.

3. Identify a suppressed subgroup and the related reported subgroup(s).

4. Repeat steps 1 and 2 for the related reported subgroup(s) to yield the number of students in each category of the outcome measure in the reported subgroup.

5. Subtract the number of students in each category of the outcome measure for the reported subgroup from the overall count for that outcome category to yield the number of students in each category of the outcome measure for the suppressed subgroup.

6. If there are more than 2 subgroups for one disaggregation (e.g., race/ethnicity), compute the counts across the categories of the outcome measure for each reported subgroup, sum subgroup counts for the reported subgroups across each outcome category, and then subtract from the overall number for that category of the outcome measure to yield the number of students in each category of the outcome measure for the suppressed subgroup(s).

All students are in one of two subgroups when student outcome measures are reported by gender, economic status, English proficiency status, migrant status, or disability status. When the data for one of the two subgroups are suppressed and the data for the other subgroup and the total are published, the suppressed data can be fully recovered. When student outcome measures are reported for race and ethnicity, subgroup data are frequently suppressed for more than one subgroup. However, the difference between the counts computed for the outcome categories of students overall and the summation across the outcome categories for the reported subgroups can be used to recover data for the suppressed subgroup(s). This recovery may yield identifying information about the students in the reporting subgroup(s) with suppressed data.

The recovery of suppressed results does not always pose a serious threat to students' personally identifiable information, but in some instances it does—the risk of identifying an individual student is a function of the distribution of students across the recovered categories.

Example 3: Suppressing Outcomes but Reporting Counts for Subgroups

The reported data in table 3 show that among 82 students who were assessed in third-grade reading, 7.3 percent (6 students) scored at the Below Basic achievement level, 42.7 percent (35 students) scored at the Basic level, 37.8 percent (31 students) scored at the Proficient level, and 12.2 percent (10 students) scored at the Advanced level. Seventy-five of the 82 students did not have an individualized education plan, and the reported data show that 8.0 percent (6 students) in this reporting subgroup scored at the Below Basic level, 42.7 percent (32 students) scored at the Basic level, 36.0 percent (27 students) scored at the Proficient level, and 13.3 percent (10 students) scored at the Advanced level.

Although the data were suppressed for students with an individualized education plan, the recovered data show that 7 of the 82 students assessed in third-grade reading were in this suppressed reporting subgroup. Further, a comparison of the overall assessment results with those for the 75 students without an individualized education plan shows that 3 of the 7 students with an individualized education plan scored at the Basic level and 4 scored at the Proficient level. These data do not provide the information needed to identify which students with an individualized education plan scored at the Proficient level and which did not. Thus, this table does not disclose an individual student's performance; however it does reveal the fact that no student with an individualized education plan scored at the Advanced level or at the Below Basic level.

In contrast, the recovered data for 8 low-income students show that 3 of these students scored at the Below Basic achievement level and 5 scored at the Basic achievement level. Thus, all students identified as low-income scored below the Proficient achievement level. If an individual student is known to be from a low-income family, the information in this table discloses that student's score as below Proficient.

The recovered data for 8 students receiving English language instruction show that 3 scored at the Below Basic achievement level, 4 scored at the Basic achievement level, and 1 scored at the Proficient level. Since parents receive their child's score along with the school report, the parents of the child who scored at the Proficient level could use the information in the published table for their child's grade to learn that each of their child's peers who received English language instruction failed to score at the Proficient achievement level.

Table 3. School-level grade 3 reading assessment results for a state with a minimum reporting size of 10

		Tested	Below Basic	Basic	Proficient	Advanced
Total	%	100	7.3	42.7	37.8	12.2
	N	82	6	35	31	10
Individualized education plan	%	100	0.0	42.9	57.1	0.0
	N	7	0	3	4	0
No individualized education plan	%	100	8.0	42.7	36.0	13.3
	N	75	6	32	27	10
English language learner	%	100	37.5	50.0	12.5	0.0
	N	8	3	4	1	0
Not English language learner	%	100	4.1	41.9	40.5	13.5
	N	74	3	31	30	10
Low income	%	100	37.5	62.5	0.0	0.0
	N	8	3	5	0	0
Not low income	%	100	4.1	40.5	41.9	13.5
	N	74	3	30	31	10

NOTE: Details may not sum to totals because of rounding.

Counts for the overall group

Some states report the percentage distribution across achievement levels for the overall population in a grade and subject along with the percentage distributions for each subgroup, but only publish the number of students tested overall for that grade and subject. This seems like it would provide more protection to students' personally identifiable information, since the number of students in each subgroup is not published. However, in many cases—especially at the school or district level for the data reported by grade and subject—there is only one unique mathematical solution that could yield the reported subgroup percentage distributions for the reported number of students overall.

Example 4: Suppressing Outcomes but Reporting Counts for Groups

In this school, 46 students were assessed in third-grade reading (table 4), and this number is known. Note that the shaded cells in the table display the data that were recovered from the reported information. Multiplying the proportions from the percentage distribution times the number in the overall group (46) shows that the 6.5 percent who scored at the Below Basic level represents 3 students (i.e., 0.65 × 46=3). The data reported by gender show that the 3 students who scored at the Below Basic level are all males. Thus, by dividing 8.3 by 3, the data show that each male student represents 2.77 percent of the number of males in the subgroup. Dividing each of the remaining percentages by 2.77 shows that there are 10 males who scored at the Basic level, 20 who scored at the Proficient level, and 3 who scored at the Advanced level.

Next, the number of males at each achievement level is subtracted from the number of students at that achievement level to recover the suppressed data for females. These calculations show that there are no females at the Below Basic level, no females at the Basic level, 7 females at the Proficient level, and 3 females at the Advanced level. The recovered data do not reveal which females scored at each of these two levels. However, when the focus of the reporting or interpretation of the data shifts to performance at or above versus below the Proficient level, the data for students scoring at the Below Basic and Basic level are combined to show the percent of students who scored below the Proficient level and the percent of students who score at the Proficient and Advanced levels are combined to show the percent of students who scored at the Proficient level. In this example, the recovered data show that all of the third-grade females in this school scored at the Proficient level or above in reading. This then discloses information about the reading achievement level of each of the third-grade females in this school.

Table 4. School-level grade 3 reading assessment results for a state with a minimum reporting size of 10

		Tested	Below Basic	Basic	Proficient	Advanced
Total	%	100	6.5	21.7	58.7	13.0
	N	46	3	10	27	6
Male	%	100	8.3	27.8	55.6	8.3
	N	36	3	10	20	3
Female	%	100	0.0	0.0	70.0	30.0
	N	10	0	0	7	3

NOTE: Details may not sum to totals because of rounding.

Counts for the overall group and subgroups reported as ranges

Another reporting approach recognizes the problem with reporting exact population counts for students assessed and, instead, reports the counts in ranges (i.e., as a categorical variable). With this approach, the percentage distribution is reported for each grade and subject overall and for each of the reporting subgroups that do not require suppression; then, instead of reporting the exact number of students in each group or subgroup, a range that includes the exact number is all that is reported for the count (e.g., instead of reporting 33 students, the number is reported as 30–39). As with the last approach, this would seem to provide more protection to students' personally identifiable information, since the exact number of students is not published. However, the range of possible values for the number of students can be used to identify the number of students that, when applied to the proportion of students at each achievement level, yields estimates that are the closest to whole numbers. Once these counts are established for the overall group and for a reported subgroup, the suppressed counts for a related subgroup can be recovered.

Example 5: Suppressing Outcomes but Reporting Ranges for Counts

The number of third-graders assessed in reading was reported as 40–49 (table 5). The percentage distribution of third-graders overall, across the achievement levels, was reported with 2 decimal places. The percentage distribution across the achievement levels was reported for the 30–39 students who did not have an individualized education plan, but the achievement results were suppressed for the 6–9 students who had one. First, the proportions from the distribution across the achievement levels were applied to each of the 10 numbers in the 40 to 49 range. The number that resulted in estimates that were closest to whole numbers is 41. This showed that, overall, 2 students scored at the Below Basic level, 5 scored at the Basic level, 15 scored at the Proficient level, and 19 scored at the Advanced level. Next, this set of steps was repeated for the 10 numbers in the 30–39 range, using the proportions from the percentage distribution across the achievement levels for students who did not have an individualized education plan. This showed that there were 34 students in this group, with none at the Below Basic level, none at the Basic level, 15 at the Proficient level, and 19 at the Advanced level.

Finally, the counts for students who did not have an individualized education plan were subtracted from the overall counts to recover the suppressed number for the students with an individualized education plan—there were 7 students in this group. Within this group, 2 scored at the Below Basic level, 5 scored at the Basic level, none scored at the Proficient level, and none scored at the Advanced level. These counts can then be used to compute the suppressed percentage distribution. The recovered data show that each of the 7 third-graders with individualized education plans scored below the Proficient level in reading. This is a disclosure of the reading achievement-level information for these 7 students

Table 5. School-level grade 3 reading assessment results for a state with a minimum reporting size of 10 and counts reported as ranges

		Percent assessed	Number tested	Below Basic	Basic	Proficient	Advanced
Total	%	100	†	4.88	12.20	36.59	46.34
	N	40–49	41	2	5	15	19
Individualized education plan	%	100	†	28.57	71.43	0.00	0.00
	N	6–9	7	2	5	0	0
No individualized education plan	%	100	†	0.00	0.00	44.12	55.88
	N	30–39	34	0	0	15	19

† Not applicable.
NOTE: Details may not sum to totals because of rounding.

Best Practices: Practices that Mitigate Disclosure Risk

The review of each state's online reporting of assessment results for schools uncovered three approaches that can help in protecting against the release of information needed to recover personally identifiable information. The first such approach involves not reporting any of the enrollment data that were used to compute the percentage distributions across the achievement-level results. The second approach starts with the first approach (i.e., the underlying enrollment counts are not reported) and collapses across outcome categories to further limit the amount of detail published. This increases the number of students included in each reported outcome category. The third approach involves suppressing subgroups other than the subgroups with less than the minimum reporting size in order to prevent the recovery of the suppressed results for the small subgroups.

Additional practices that support public reporting while protecting student privacy were identified and are discussed in this section. The first involves the reporting of background data on enrollment by grade and enrollment by student characteristics for a school or district. The second involves protecting data at the ends of the distribution, or at the low and high values for a rate, to avoid reporting that a small number of students (or nearly all students) have a specific outcome.

Each of these practices taken alone does not necessarily address each of the potential sources of disclosure, but they do reflect practices that, when taken in combination, may lead to improved protection of personally identifiable information about individual students in published tables.

No Counts Published

Eight states were identified that publish student assessment results by grade and subject for the overall student population and for the reportable subgroups (i.e., those subgroups that do not require suppression) only as a percentage distribution across the achievement levels. In these states, the school reports do not include counts of the number of students assessed overall or of the number of students assessed in each of the reporting subgroups. However, since too much precision in the percentages can limit the possible options for the underlying counts, limiting the percentages reported to whole numbers increases the number of possible options for the underlying counts. This helps protect the suppressed data for small groups. It also helps protect the counts for small categories within outcome measures for the reported subgroups. The following example of school-level third-grade reading results shows that while the relative relationships across achievement levels within and across subgroups are evident, the absence of the counts used to compute the percentage distributions prevents the recovery of the suppressed data.

Example 6: Best Practices: No Counts Published

Table 6 shows assessment results only as percentage distributions reported as whole numbers. This, coupled with the fact that no counts are reported, protects the suppressed data from disclosure (table 6). The table shows that 13 percent of the students scored at the Below Basic level, 44 percent scored at the Basic level, 27 percent scored at the Proficient level, and 16 percent scored at the Advanced level. Relatively more male than female students and more low-socioeconomic status than non-low-socioeconomic status students performed at the Below Basic level. The data are suppressed for the English language learner subgroup because there are fewer than 10 students in the subgroup.

Table 6. Percentage distribution of school-level grade 3 reading assessment results in a state with a minimum reporting size of 10 and no counts

	Below Basic	Basic	Proficient	Advanced
Total	13	44	27	16
Male	17	47	23	13
Female	9	42	30	18
Low SES	28	39	22	11
Not low SES	7	47	29	18
English language learner	*	*	*	*
Not English language learner	6	44	31	19

* Not reported to protect subgroups with fewer than 10 students.
NOTE: Details may not sum to totals because of rounding. SES = Socioeconomic status.

Collapsing Across Outcome Categories

Seven states limited their reporting of achievement results to two categories—those at or above the level established by the state for successful performance and those who did not score in the successful range. Collapsing across outcome categories is useful when there are a small number of students in one or more of the outcome categories. This approach, combined with the decision to not report the underlying counts, is another way of increasing the protection of student privacy in reported summary tables.

Example 7: Best Practices: Collapsing across Outcome Categories

Collapsing across outcome categories and displaying the assessment results only as a percentage distribution protects the underlying counts from disclosure. Collapsing the data used in the previous example, 57 percent of the students scored at or below the Basic level, and 43 percent scored at or above the Proficient level (table 7). Relatively more male then female students (64 percent versus 51 percent) and low socioeconomic status than not low socioeconomic status students (67 percent versus 53 percent) scored at the Below Basic level. The data are suppressed for the English language learner subgroup because there are less than 10 students in the subgroup.

Table 7. Percentage distribution of school level, grade 3 reading assessment results collapsed in a state with a minimum reporting size of 10 and no counts

	Basic or below	Proficient or above
Total	57	43
Male	64	36
Female	51	48
Low SES	67	33
Not low SES	53	47
English language learner	*	*
Not English language learner	50	50

* Not reported to protect subgroups with fewer than 10 students.
NOTE: Details may not sum to totals because of rounding. SES = Socioeconomic status.

Counts Published with Additional Suppression

One state provides counts for the overall number of students assessed in a specific grade and subject and for students in reportable subgroups. However, instead of suppressing only the subgroups that do not meet the minimum reporting size, subgroups related to the suppressed group are also suppressed. This is referred to as "complementary suppression." That is, a subgroup with less than 10 students is suppressed, and one (or more) of the other subgroups that combine with the small subgroup to account for a larger share of the students in the overall group is also suppressed. The following example of school-level third-grade reading results provides an illustration of this approach.

Example 8: Best Practices: Schools Counts Published with Additional Suppression

This example includes two schools. The school-level report is designed to display results by gender, race and ethnicity, low-income status, and individualized education plan status. School 1, with 30 students, had a number of reporting subgroups with fewer than 10 students. Suppressing the assessment results for the small subgroups and suppressing the outcome measure for a related category (i.e., complementary suppression of additional rows of the table) protects the reported data at the school level, but leads to the loss of information. As shown in table 8, data were suppressed for the 27 White students because there were fewer than 10 students in each of the other racial and ethnic subgroups (i.e., 2 Native American students and 1 Black student). Data were suppressed for the 21 low income students because there were fewer than 10 students who were not low income. Data were also suppressed for the 21 students without an individualized education plan, because only 9 students had individualized education plans. By comparison, assessment data were reported for the 30 third-grade students overall, and for the 12 male and 18 female students because the minimum reporting threshold of 10 students was exceeded in each case.

Table 8. School 1: Number tested and percentage distribution of grade 3 reading assessment results with a minimum reporting size of 10 and complementary row suppression

	Number tested	Below Basic	Basic	Proficient	Advanced
Total	30	16.7	56.7	20.0	6.7
Male	12	25.0	58.3	16.7	0.0
Female	18	11.1	55.6	22.2	11.1
White	27	*	*	*	*
Native American	2	*	*	*	*
Black	1	*	*	*	*
Low income	21	*	*	*	*
Not low income	9	*	*	*	*
Individualized education plan	9	*	*	*	*
No individualized education plan	21	*	*	*	*

* Not reported to protect subgroups with fewer than 10 students.

School 2, with 45 students, had 10 or more students in each reporting group. As a result, no data were suppressed and the third-grade reading assessment results were reported for each of the reporting variables—gender, race and ethnicity, low income status, and individualized education plan status (table 9).

Table 9. School 2: Number tested and percentage distribution of grade 3 reading assessment results with a minimum reporting size of 10 and complementary row suppression

	Number tested	Below Basic	Basic	Proficient	Advanced
Total	45	2.2	22.2	62.2	13.3
Male	18	5.6	27.8	55.6	11.1
Female	27	0.0	18.5	66.7	14.8
White	20	0.0	10.0	65.0	25.0
Native American	10	10.0	40.0	50.0	0.0
Black	15	0.0	26.7	66.7	6.7
Low income	14	7.1	21.4	64.3	7.1
Not low income	31	0.0	22.6	61.3	16.1
Individualized education plan	11	9.1	72.7	18.2	0.0
No individualized education plan	34	0.0	5.9	76.5	17.6

These two schools are the only schools in a district that include the third grade. When the data for the two schools were combined at the district level, there were 10 or more students in each reporting group. The resulting data are displayed in the next example.

Example 9: Best Practices: District Counts Published with Additional Suppression

Since there were more than 10 students in each reporting subgroup at the district level, the district table based on the schools in example 8 (tables 8 and 9) was produced with full details reported for each reporting group. Table 10 displays these results.

Table 10. Number tested and percentage distribution of district-level grade 3 reading assessment results with a minimum reporting size of 10 and complementary row suppression

	Number tested	Below Basic	Basic	Proficient	Advanced
Total	75	8.0	36.0	45.3	10.6
Male	30	13.4	40.0	40.0	6.7
Female	45	4.4	33.3	48.9	13.3
White	47	6.4	38.3	40.4	14.9
Native American	12	16.7	41.7	41.7	0.0
Black	16	6.3	25.9	62.5	6.3
Low income	35	17.1	54.3	25.7	2.8
Not low income	40	0.0	20.0	62.5	17.5
Individualized education plan	20	30.0	55.0	15.0	0.0
No individualized education plan	55	0.0	29.1	56.4	14.5

But with all of the details published for school 2 and for the district, the percentage distribution across the achievement levels in each row can be converted to proportions. The proportions can then be applied to the number of students in the reporting subgroup to compute the number of students at each achievement level in each reporting group. Once this is done at the district level and for school 2, all of the suppressed data for school 1 can be recovered. For example, 38.3 percent of the 47 White third graders in the district scored at the Basic achievement level. Multiplying 0.383 times 47 shows that 18 White third graders in the district scored at the Basic achievement level. The results for White third graders in school 2 show that 10 percent of the 20 students in this subgroup scored at the Basic achievement level. Multiplying 0.10 times 20 shows that 2 White third graders in School 2 scored at the Basic achievement level. Subtracting the 2 students from School 2 from the 18 students in the district reveals the fact that there were 16 White third graders in School 1 who scored at the Basic achievement level. These 16 students comprise 59.3 percent of the 27 White third graders in school 1. These procedures were repeated to recover each of the percentages that were suppressed for school 1 in table 8. The recovered results for school 1 are shown in the shaded cells in table 11 which show that the 2 Native American third graders scored at or below Basic, the 1 Black third grader scored below Basic, and 23.8 percent of the 21 low income students scored below Basic and the other 76.2 percent scored at the Basic level. When the results for students who scored at the below Basic and Basic levels are combined to show the percent who scored below proficient, the data show disclosures of the fact that all students who were Native American, Black, or low income scored below the Proficient level. Furthermore, the parents of the 1 third grade student in school 1 with an individualized education plan who scored at the Proficient achievement level (i.e., 11.1 percent of 9 students is 1 student) know that the other third graders with individualized education plans each failed to reach the Proficient achievement level.

Table 11. School 1: Number tested and percentage distribution of grade 3 reading assessment results with suppressed percents recovered

	Number tested	Below Basic	Basic	Proficient	Advanced
Total	30	16.7	56.7	20.0	6.7
Male	12	25.0	58.3	16.7	0.0
Female	18	11.1	55.6	22.2	11.1
White	27	11.1	59.3	22.2	7.4
Native American	2	50.0	50.0	0.0	0.0
Black	1	100.0	0.0	0.0	0.0
Low income	21	23.8	76.2	0.0	0.0
Not low income	9	0.0	11.1	66.7	22.2
Individualized education plan	9	55.6	33.3	11.1	0.0
No individualized education plan	21	0.0	66.7	23.8	9.5

* Not reported to protect subgroups with fewer than 10 students.

This example illustrates the fact that it is not enough to simply suppress results at the school level, since comparisons of data published for other schools and the district can be used to recover suppressed results within a school. To avoid the recovery of suppressed school level results, the results for other schools in the district and the results for the district must also be taken into account. If the results for a specific subgroup are suppressed in at least two schools, the suppressed results for each school cannot be recovered from the results reported for other schools and the district. However, when the results are suppressed for a specific subgroup in only one school, to protect the suppressed results from recovery, the results for that subgroup must be suppressed for either another school in the district or for the district.

To protect results that are suppressed at the district level, the same precautions must be taken across district and state results. To protect suppressed results from recovery, if the results are suppressed for a specific subgroup in one district, the results

for that subgroup must be suppressed for a second district in the state.

It is important to note that this problem is not limited to applications that use complementary suppression across related subgroups. The same comparisons between district results and the results reported for other schools in the district or between state results and the results reported for other districts in the state can be applied when the results are suppressed for a single subgroup (i.e., without complementary subgroup suppression). Care must be taken to ensure that the suppressed results for a subgroup in a single school or single district cannot be recovered using reported data for other schools in the district or other districts in the state. This can be achieved by ensuring that the results for a suppressed subgroup are suppressed in two schools. Alternatively, in districts with only one school for a grade, the results for the suppressed subgroup must also be suppressed at the district level. Similarly, the results for a suppressed subgroup must be suppressed for two districts in a state.

Reporting School-, District-, or State-Level Background Information

In reports of outcome measures, some school-, district-, or state-level reports display background information on the distribution of students in a school, district, or state in two separate summary tables. One summary table reports the total number of students enrolled and the percentage of students enrolled by grade. The second summary table reports the total number of students enrolled and the percentage of students in each of the reporting subgroups (e.g., gender, race and ethnicity, English proficiency status, migrant status, disability status, and economic status). Thus, rather than providing the exact number or percentage of students in each grade in each reporting subgroup, the report gives a portrait of the school, district, or state. However, if the number of students reported for an individual grade is the same as the number of students enrolled on the assessment date, that number, along with the report of the percentage of the students who participated in the assessment, can be used with the percentage distribution across the achievement levels to recover the underlying numbers of students who scored at each achievement level.

Three things can be done to counter this problem. First, use background enrollment counts for a day other than that of the assessment administration and clearly label the date of the background enrollment counts and the date of the assessment in public reports to establish the fact that they are different. Second, report the percentage distribution for the background data and for the results reported across the achievement levels only in whole numbers. This decreases the precision of the reported percentages, which lowers the chance of an accurate recovery of the numbers of students in both reported and suppressed results. Third, report the percentage of students assessed as a whole number.

Example 10: Best Practices: Reporting Background Information

Table 12 provides an example of school-level data for enrollment by grade for an elementary school with grades K–6. The shaded cells are not included in the reported table, but are included here to illustrate the added protection from reporting the percentage distribution without any decimal places. For example, 4 of the 7 grades are reported as being 14 percent of the school's enrollment; the underlying data show that the more precise percentages are 13.9, 14.5, 13.6, and 14.2. The state assessment in this state is administered in March of each school year; reporting enrollment data from 5 months earlier in the school year is likely to result in some differences from the enrollment data at the time of the assessment.

Table 13 displays school-level enrollment data reported by student characteristics for the same elementary school. Again, the patterned cells are not included in the reported table. Taken together, these tables provide a profile of the school without providing the level of detail needed to recover the underlying counts for the outcome measures reported for the school.

Table 12. Elementary school enrollment, by grade

	Number	Unrounded percent	Percent
Total	359	100.0	100
Kindergarten	50	13.9	14
Grade 1	52	14.5	14
Grade 2	54	15.0	15
Grade 3	49	13.6	14
Grade 4	48	13.4	13
Grade 5	51	14.2	14
Grade 6	55	15.3	15

Table 13. Elementary school enrollment, by selected characteristics

	Number	Unrounded percent	Percent
Total	359	†	†
Male	185	51.5	52
Female	174	48.5	48
White	221	61.6	62
Black	70	19.5	19
Hispanic	59	16.4	16
Asian	*	*	*
Native American	*	*	*
Low income	100	27.9	28
Not low income	259	72.1	72
Individualized education plan	59	16.4	16
No individualized education plan	300	83.6	84
English language learner	40	11.1	11
Not English language learner	319	88.9	89

† Not applicable.
* Not reported to protect subgroups with fewer than 10 students.

Recoding the Ends of the Distribution

Another protection implemented by a number of states involves bottom or top coding the results at the tails of the percentage distribution, or for high and low rates. This is typically done by coding all percentages above 95 percent as greater than 95 percent and coding all percentages below 5 percent as less than 5 percent. This is done to avoid reporting the fact that all, or nearly all, of the students in a reporting subgroup share the same achievement level or the same outcome or that very few or none of the students have a particular outcome.

Ideally, this approach is intended to protect categories with 0 to 2 fewer than all students in a reporting category or, conversely, categories with 0 to 2 students. However, with reporting subgroups of 10 to 19 students, all of the percentages of 10 percent or less are based on only 1 student (e.g., 1 of 19 students is 5 percent and 1 of 10 students is 10 percent, while 2 of 19 is 11 percent and 2 of 10 is 20 percent). As a result, with reporting subgroups of 10 to 19 students, even reporting a category as 10 percent or less is no different than reporting that there is at most only 1 student in the category.

The extent of recoding required to protect small categories is related to the size of the subgroup, with a larger recoded range required for smaller subgroups. At a minimum, results should not be published for outcomes based on the experiences of 1 student. The goal is to ensure that each recoded percent could include at least 2 students. Additional protection is provided by including counts of students in the range of recoded percentages where the recoded percent could include at least 3 students (i.e., the threshold rule of 3). For example, in reporting outcome measures for subgroups of 10 to 20, recoding the ends of the distribution to 20 percent or less and 80 percent or more would result in recoding all percentages for categories based on 0 to 2 students (i.e., 20 percent of 10 is 2).[6] In addition, categories of 3 students would be included in the recoded category when there are 15 or more students in the subgroup (i.e., 3 out of 15 is 20 percent).

In reporting outcome measures for groups of 21 to 40, recoding the ends of the distribution to 10 percent or less and 90 percent or more would result in recoding all percentages based on categories of 0 to 2 students. In this recode, categories of 3 students would be included in the recoded category when there are 30 or more students in the subgroup (i.e., 3 out of 30 is 10 percent).

When there are 41 to 100 students, recoding the ends of the distribution to 5 percent or less and 95 percent or more ensures results based on 0 to 2 students when there are 41 students and 0 to 4 students when there are 100 students (above 59 students, this recode would include categories of 3 students). Similarly, for groups of 101 to 300 students, recoding the ends of the distribution to 2 percent or less and 98 percent or more ensures reporting results based on 0 to 2 students when there are 101 students and 0 to 6 students when there are 300 students (above 149 students this recode includes categories of 3 students) . Finally, for groups of more than 300 students, recoding the ends of the distribution to 1 percent or less and 99 percent or more ensures results based on 0 to 3 students at a minimum

Recoding the percentages at one end of a percentage distribution is not necessarily enough to protect the original contents of the recoded category, since the sum of the reported categories subtracted from 100 percent yields the percent that was recoded.

To protect the recoded categories, additional recoding is needed. For groups of 10 to 20 students, the results should be collapsed into two categories and percentages between 21 and 79 should be reported in 10 percentage point ranges. For groups of 21 to 40 students, the percentages in categories of an outcome measure should be recoded in 10 percentage point ranges. For groups of 41 to 200 students, the percentages in categories of an outcome measure should be recoded in 5 percentage point ranges. For groups of 201 or more students, reporting the percentages in categories of an outcome measure as whole numbers provides sufficient recoding (i.e. there are at least 2 counts that could yield each reported percent).

To further protect small categories, if one subgroup includes 200 or fewer students, any related subgroups (i.e., those that combine to sum to the total) with more than 200 students should be recoded using the ranges for 200 students.

[6] Reporting results based on fewer than 10 students while ensuring that there could be at least 2 students in a reported category requires more extensive top and bottom coding and would limit the number of reportable outcomes to a small enough set of possible outcomes that they would not be well protected. For example, with results based on 6 students, 2 students account for 33 percent, and recodes of 33 and 67 percent leave only 1 response option that could be reported. Similarly, with 7 students, the recodes would be 29 and 71 percent, leaving 2 response options for reporting; with 8 students, the recodes would be 25 and 75 percent, leaving 3 response options for reporting; and with 9 students, the recodes would be 22 and 78 percent, leaving only 5 response options for reporting.

Example 11: Best Practices: Recoding the Distribution

Table 14 in this example shows the number of students and the actual and recoded percentage distributions for the school-level third-grade reading assessment results for 32 students for this reporting option. The shaded cells are not publicly reported. Table 14 displays the data with reporting subgroups less than 10 suppressed and the categories of other subgroups recoded to protect small categories. For the overall results of the 32 students, each category is recoded into a 10 percentage point range to protect small categories in the subgroups in the table. Given that there are only 10 students in the Hispanic subgroup, the 0 in the Advanced category is combined with the 10 percent in the proficient category and recoded to less than or equal to (≤) 20 percent at or above proficient, and the 50 percent at the Basic level is combined with the 40 percent at the Below Basic level and recoded to greater than or equal to 80 percent. The data for the 22 White students are recoded, with the 0 percent in the Below Basic category recoded to less than or equal to 10 percent and the other three categories recoded into 10 percentage point ranges. Since there are fewer than 10 students with individualized education plans, the data for this subgroup and the data for students who do not have individualized education plan are suppressed. The outcome measures for the 12 English language learners and the subgroup of 20 students who are not English language learners are reported for those students scoring at the proficient or above level and those performing at or Below the Basic level.

Table 15 follows the same format and shows the results for the district-level third-grade reading assessment results. With 320 students in the group, the results for the 3 students in the advanced category that account for 1 percent of the total are recoded to less than or equal to (≤) 1 percent, and the other three categories are reported as percentages that are rounded to whole numbers. With 198 White students and 122 Hispanic students, the results for the 3 Advanced students in the White subgroup and for 0 Advanced students in the Hispanic subgroup are both recoded to less than or equal to (≤) 2 percent, and the other three categories in each subgroup are recoded into 5 percentage point ranges. With 40 students with individualized education plans, the Advanced category for these students is recoded to less than or equal to (≤) 10 percent, and the remaining categories are recoded into 10 percentage point ranges. The data for the 280 students in the related subgroup who do not have individualized education plans are recoded following the procedures that apply to 200 students, with the 1 percent at the Advanced level recoded to less than or equal to (≤) 2 percent and the other three categories recoded into 5 percentage point ranges. Finally, because there are only 12 students who are English language learners, the Advanced category for these students is combined with the Proficient category and reported as 21 to 29 percent, and the Below Basic and Basic categories are combined and reported as 70 to 79 percent. The data for the 308 students in the related subgroup who are not English language learners are recoded, with the percent at the Advanced level reported as less than or equal to (≤) 2 percent and the other three categories recoded into 5 percentage point ranges.

Table 14. School-level grade 3 reading assessment results for a state with a minimum reporting size of 10

		Percent assessed	Tested	Below Basic	Basic	Proficient	Advanced	
Total	N	†	32	4	10	11	7	
	%	100	100	13	31	34	22	Actual
	%	100	100	11–19	30–39	30–39	20–29	Reported
White	N	†	22	0	5	10	7	
	%	100	100	0	23	45	32	Actual
	%	100	100	≤10	21–29	40–49	30–39	Reported
Hispanic	N	†	10	4	5	1	0	
	%	100	100	40	50	10	0	Actual
	%	100	100	†	≥80	≤20	†	Reported
Individualized education plan	N	†	7	4	3	0	0	
	%	100	*	*	*	*	*	<10
	%	100	*	*	*	*	*	<10
No individualized education plan	N	†	25	0	7	11	7	
	%	100	100	0	28	44	28	Actual
	%	100	*	*	*	*	*	Suppressed
English language learner	N	†	12	4	5	2	1	
	%	100	100	33	42	17	8	Actual
	%	100	100	†	70–79	21–29	†	Reported
Not English language learner	N	†	20	0	5	9	6	
	%	100	100	0	25	45	30	Actual
	%	100	100	†	21–29	70–79	†	Reported

† Not applicable.
* Not reported to protect subgroups with fewer than 10 students.
NOTE: Details may not sum to totals because of rounding and recoding.

Table 15. District level, Grade 3 reading assessment results for a state with a minimum reporting size of 10

		Percent assessed	Tested	Below Basic	Basic	Proficient	Advanced	
Total	N		320	40	167	110	3	
	%	100	†	13	52	34	1	Actual
	%	100	†	13	52	34	≤1	Reported
White	N		198	0	105	90	3	
	%	100	†	0	53	45	2	Actual
	%	100	†	≤2	50–54	45–49	≤2	Reported
Hispanic	N		122	40	62	20	0	
	%	100	†	33	51	16	0	Actual
	%	100	†	25–29	50–54	15–19	≤2	Reported
Individualized education plan	N		40	25	15	0	0	
	%	100	†	63	38	0	0	Actual
	%	100	†	60–69	30–39	≤10	≤10	Reported
No individualized education plan	N		280	15	152	110	3	
	%	100	†	5	54	39	1	Actual
	%	100	†	5–9	50–54	35–39	≤2	Reported
English language learner	N		12	4	5	2	1	
	%	100	†	33	42	17	8	Actual
	%	100	†	†	70–79	21–29	†	Reported
Not English language learner	N		308	36	162	108	2	
	%	100	†	12	53	35	1	Actual
	%	100	†	10–14	50–54	35–39	≤2	Reported

† Not applicable.
NOTE: Details may not sum to totals because of rounding and recoding.

Recommendations

This review and analysis of current reporting practices illustrates that some practices work better than others in protecting suppressed results and, thus, in protecting against disclosures of personally identifiable information about individual students. It is important to note that each of the practices requires some loss of information. The challenge rests in identifying practices that protect information about individual students while minimizing the negative impact on the utility of the publicly reported data. Drawing upon the review and analysis presented in this brief leads to recommended reporting rules to be used in producing reports of percentages and rates to describe student outcomes to the public. These rules are intended for use in the public release of new data.

Rules 1 through 4 and 6 and 7 are general reporting rules. Rule 5 is guided by the number of students in the reporting group or subgroups; the underlying principle is that the amount of detail that can be reported while protecting each student's privacy is related to the number of students in a reporting group or subgroup—that is, more detail can be reported for larger groups. Rule 5a applies to instances in which there are more than 300 students in each of a set of related reporting subgroups (e.g., in each race/ethnicity group, for students with and without an individualized education plan, for students receiving or not receiving instruction as an English language learner). Rule 5b applies to instances in which the smallest reporting subgroup within a set of related reporting subgroups has 201 to 300 students. Rule 5c applies to instances in which the smallest reporting subgroup within a set of related reporting subgroups has 101 to 200 students. Rule 5d applies when the smallest reporting subgroup in a set of related subgroups has 41 to 100 students. Rule 5e applies when the smallest reporting subgroup in a set of related subgroups has 21 to 40 students. Rule5f applies when the smallest reporting subgroup in a set of related subgroups has 10 to 20 students.

Reporting Rules

1. Minimize the amount of enrollment details reported in the profile of the school, district, or state in reports of outcome measure results. If possible, use enrollment data for a different date than that of the reported outcome measures and label the different dates (e.g., report enrollment data for a date different from the assessment date, such as fall enrollment for a spring assessment). In so doing, tell the readers that the data on student enrollment by grade and by selected student characteristics are included to provide context for the results presented but should not be assumed to exactly match the student composition at the time the outcome was measured.
    a. Report the percentage distribution of students by grade at the school, district, or state level in a standalone table without any of the outcome measures or reporting subgroup details.
    b. Report the percentage distribution of students by reporting subgroup at the school, district, or state level in a standalone table without any of the outcome measures or enrollment by grade details.
    c. Do not report the details of the enrollment data within each reporting subgroup by individual grades.
4. Use a minimum of 10 students for the reporting subgroup size limitation.
    a. Suppress results for all reporting groups with 0 to 9 students.
    b. Suppress results for reporting subgroups with 0 to 9 students and suppress each of the related reporting subgroups regardless of the number of students in the subgroup (i.e., suppress the other subgroup(s) of the set of subgroups that sum to the overall group). In instances with 3 or more subgroups, the subgroups with 0 to 9 students can be combined with each other or with the smallest reportable subgroup to form an aggregated subgroup of 10 or more students to allow for the reporting of data for larger subgroups.
3. Use only whole numbers when reporting the percentage of students for each category of an outcome measure (e.g., the percentage assessed).

4. Do not report the underlying counts for the subgroup or group totals (i.e., the denominators of the percentages); also do not report the underlying counts of students in individual outcome categories (i.e., the numerators).

5. **To implement the next step in the data protection procedure in the remaining reporting groups and subgroups, the approach used is determined by the number of students in the smallest reporting subgroup among a set of related groups or subgroups (i.e., groups that in combination sum to the total). To protect student privacy:**

    a. **For reporting variables/outcome measures with more than 300 students and no related subgroup with fewer than 200 students, use the following approach:**

        i. Recode categories with values of 99 to 100 percent to greater than or equal to 99 percent (≥ 99 percent).

        ii. Recode categories with values of 0 to 1 percent to less than or equal to 1 percent (≤ 1 percent).

        iii. Otherwise, report the percentage of students in each category using whole numbers.

    b. **For reporting variables/outcome measures with 201 to 300 students and no related subgroup with fewer than 200 students, use the following approach:**

        i. Recode categories with values of 98 to 100 percent to greater than or equal to 98 percent (≥ 98 percent).

        ii. Recode categories with values of 0 to 2 percent to less than or equal to 2 percent (≤ 2 percent).

        iii. Otherwise, report the percentage of students in each category using whole numbers.

    c. **For reporting variables/outcome measures in which the number of students ranges from 101 to 200,** use the following option in this group and all related subgroups with more than 200 students:

        i. Recode categories with values of 98 to 100 percent to greater than or equal to 98 percent (≥ 98 percent).

        ii. Recode categories with values of 0 to 2 percent to less than or equal to 20 percent (≤ 2 percent).

        iii. Recode the percentage in each remaining category in all reporting groups or subgroups to intervals as follows (3–4, 5–9, 10–14, 15–19, . . ., 85–89, 90–94, 95–97).

    d. **For reporting variables/outcome measures in which the number of students in the smallest reporting group or subgroup ranges from 41 to 100,** use the following option in that group or subgroup and use option 5c for each related reporting group or subgroup with more than 100 students:

        i. Recode categories with values of 95 to 100 percent to greater than or equal to 95 percent (≥ 95 percent).

        ii. Recode categories with values of 0 to 5 percent to less than or equal to 5 percent (≤ 5 percent).

        iii. Recode the percentage in each remaining category in all reporting groups or subgroups to intervals as follows (6–9,10–14, 15–19, 20–24, . . ., 85–89, 90–94).

e. **For reporting variables/outcome measures in which the number of students in the smallest reporting group or subgroup ranges from 21 to 40,** use the following option for that group or subgroup, use option 5d for each related reporting group or subgroup with 41 to 100 students, and use option 5c for those with more than 100 students:

   i. Recode categories with values of 90 to 100 percent to greater than or equal to 90 percent (≥ 90 percent).

   ii. Recode categories with values of 0 to 10 percent to less than or equal to 10 percent (≤ 10 percent).

   iii. Recode the percentage in each remaining category in all reporting groups or subgroups to intervals as follows (11–19, 20–29, . . ., 80–89).

f. **For reporting variables with 10 to 20 students in the smallest subgroup,** use the following option for that group or subgroup, use option 5e for each related group or subgroup with 21 to 40 students, use option 5d for those with 41 to 100 students, and use option 5c for those with more than 100 students:

   i. Collapse all outcome measures to only two categories, using the same collapsing rules across all subgroups for each outcome measure (e.g., assessment results collapsed to below the proficient level and at or above the proficient level by sex, racial and ethnic groups, disability status, etc.).

   ii. Recode categories with values of 0 to 20 percent to less than or equal to 20 percent (≤ 20 percent), and recode the other category to greater than 80 percent (> 80 percent).

   iii. If both collapsed categories have percents of 21 to 79 percent, recode the percentage in each collapsed category to intervals as follows (21–29, 30–39, . . ., 70–79).

6. For each outcome measure reported at the district level, if results for a group or subgroup have been collapsed, recoded, or suppressed in only one school in the district, apply the same collapsing, recoding, or suppression rule for that group or subgroup in a second school or at the district level (i.e., for any specific measure and group or subgroup, there must be either no school-level data suppressed for a specific subgroup or the data for that subgroup must be suppressed for at least 2 schools or for one school and the district).

7. For each outcome measure reported at the state level, if results for a group or subgroup have been collapsed, recoded, or suppressed in only one district in the state, apply the same collapsing, recoding, or suppression rule for that group or subgroup in a second district (i.e., for any specific measure and group or subgroup, there must be either no district-level data suppressed for a specific subgroup or the data for that subgroup must be recoded or suppressed for at least 2 districts).

Summary

This Brief discusses the potential for the disclosure of personally identifiable information in summary school-, district-, and state-level reports from education records using current reporting practices. Building on current best practices, the Brief outlines reporting recommendations. Primarily, the goal of these reporting recommendations is to maximize the reporting of student outcomes while protecting students' personally identifiable information.

While it would be easier to have only one set of reporting recommendations, the reporting rules are intended to maximize the amount of detail that can be safely reported without allowing the disclosure of student outcome measure categories based on small numbers of students. A secondary goal of these recommendations is to maximize uniformity in reporting practices across states in order to facilitate cross-state comparisons.

The recommendation to provide data on enrollment by grade and enrollment by student characteristics that are not identical to those for the day the outcome is measured is intended to prevent the statistical manipulation of the data to recover protected student information. However, this may not always be possible, and in some instances, these data may not change over the course of a school year. Thus, the reporting rules that are linked to the number of students included in a subgroup are intended to add additional protections by ensuring that, if the subgroup size is known, each reported category could include at least two students. Further, if the subgroup size is not known, each reported category could include at least three students.

There are multiple approaches to statistical data protection. The recommendations here were selected with the goal of maximizing the amount of information that can be released while protecting personally identifiable student information through a relatively straightforward set of rules that can be easily implemented. For those readers wanting to read further on the topic of statistical data protection, please see Duncan et. al. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics;* Willenborg and de Waal (2001) *Statistical Disclosure Control in Practice;* Federal Committee on Statistical Methodology Working Paper 22, *Report on Statistical Disclosure Limitation Methodology;* and the American Statistical Association, Committee on Privacy and Confidentiality website, *Key Terms/Definitions in Privacy and Confidentiality.*

NCES welcomes input on these recommendations.

References

American Statistical Association, Committee on Privacy and Confidentiality. *Key Terms/ Definitions in Privacy and Confidentiality.* Alexandria, VA: Retrieved from http://www.amstat.org/committees/pc/keyterms.html on 6/17/2010.

Code of Federal Regulations, Title 34—Education, Part 200. *Improving the Academic Achievement of the Disadvantaged, Section 200.7, Disaggregation of Data,* (34CFR200.7). Washington, DC: GPO Access CFR. Retrieved from http://edocket.access.gpo.gov/cfr_2010/julqtr/34cfr200.7.htm on 10/10/2010.

Code of Federal Regulations, Title 34—Education, Part 99. *Family Educational and Privacy Rights,* (34CFR99). Washington, DC: GPO Access e-CFR. Retrieved from http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid=44d350c26fb9cba4a156bf805f297c9e&tpl=/ecfrbrowse/Title34/34cfr99_main_02.tpl.

Duncan, George T., Jabine, Thomas B. and de Wolf, Virginia A., Editors. (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Panel on Confidentiality and Data Access, National Research Council. Washington, DC: National Academy Press.

Federal Register, Office of Management and Budget, *Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA),* Washington DC: Vol. 72, No. 115 / Friday, June 15, 2007. Retrieved from http://www.whitehouse.gov/sites/default/files/omb/assets/omb/fedreg/2007/061507_cipsea_guidance.pdf on 9/9/2010.

Office of Management and Budget, Federal Committee on Statistical Methodology, (2005). Statistical Policy Working Paper 22, *Report on Statistical Disclosure Limitation Methodology.*

Retrieved from http://www.fcsm.gov/workingpapers/spwp22.html on 9/9/2010.

Office of Management and Budget, OMB Memorandum M-03-22, *OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002*. Retrieved from http://www.whitehouse.gov/omb/memoranda_m03-22/ on 9/9/2010.

Office of Management and Budget, OMB Memorandum M-07-16, *Safeguarding Against and Responding to the Breach of Personally Identifiable Information*. Retrieved from http://www.whitehouse.gov/sites/default/files/omb/assets/omb/memoranda/fy2007/m07-16.pdf on 9/9/2010.

U.S. Code, Title 20—Education, Chapter 31—General Provisions Concerning Education, Subchapter III—General Requirements and Conditions Concerning Operation and Administration of Education Programs: General Authority of Secretary, Part 4—Records, Privacy, Limitation on Withholding Federal funds, Section 1232g. *Family Educational and Privacy Rights,* (20USC1232g). Washington, DC: GPO Access. Retrieved from http://frwebgate4.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=799486197532+0+1+0&WAISaction=retrieve.

U.S. Code, Title 20—Education, Chapter 70—Strengthening and Improvement of Elementary and Secondary Schools, Subchapter I—Improving the Academic Achievement of the Disadvantaged, Part A—Improving Basic Programs Operated by Local Educational Agencies, Subpart 1—Basic Program Requirements, Section 6311. *State Plans,* (20USC6311). Washington, DC: GPO Access. Retrieved from http://frwebgate2.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=bULwJH/21/1/0&WAISaction=retrieve.

U.S. Code, Title 20—Education, Chapter 76—Education Research, Statistics, Evaluation, Information, and Dissemination, Subchapter II—Educational Technical Assistance, Section 9607. *Grant Program for Statewide, Longitudinal Data Systems,* (20USC9607). Washington, DC: GPO Access. Retrieved from http://frwebgate3.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=FKr6BA/0/1/0&WAISaction=retrieve on 9/9/2010.

Public Law 110-69, America Competes Act, Title VI—Education, Section 6401. Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ069.110 on 9/10/2010.

Public Law, 111-05, American Recovery and Reinvestment Act, Title VIII—Education, Institute of Education Sciences. Washington, DC: GPO Access. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_public_laws&docid=f:publ005.111 on 9/10/2010.

Willenborg, L., and De Waal, T. (2001). *Elements of Statistical Disclosure Control,* Vol. 155, Lecture Notes in Statistics, New York, NY: Springer.

Examples from Other States' LDS Projects – Data Sharing

Here are summaries of data-sharing practices of longitudinal data system (LDS) projects in six other states as of January 2011.

Note that SLDS data sharing processes and tools are changing rapidly in many states with SLDS grants. To review up-to-date progress for any state, follow links provided on these documents to their web portals.

Arkansas
Their investments in data-sharing capabilities reflect their primary goal of "deliver data to the teacher" for outcomes improvement. They have implemented technologically sophisticated, interactive tools and training activities to deliver usable information to state education personnel at all levels, and there is relatively little emphasis on data sharing with external researchers.

Florida
Frequently cited as most advanced SLDS project, with full P-20 linking and widespread data sharing for multiple audiences and purposes, their system includes a centralized, integrated data warehouse and information portal, and formal policies and processes for data-sharing with researchers.

Kansas
They have developed a formal data-sharing policy, request and approval process implemented through a web portal hosted by a four-agency government/academic consortium. In December 2010, Kansas representatives reported they are nearing the end of their grant period and focusing on streamlining and producing a greater return on investment. Their next challenge is to increase data-driven decision-making through re-prioritization and limiting focus to policy and research stakeholders.

North Carolina
In contrast to Arkansas's system where data are primarily made available to internal state education stakeholders and not to external researchers—North Carolina has a distributed (de-centralized) SLDS system including a long-term partnership between a major academic research center and the state department of education. Often cited as a model for open but legal and secure sharing of data for research.

Pennsylvania
Recognized as a national model for integrating several early childhood development and early learning program data systems, they are beginning to link EL data with K12 and post-secondary data systems for a full P-20 system.

Texas
Like Florida, Texas has at least a decade of experience in developing their education data systems. In 2008, they launched a major upgrade of their aging system technology, evidenced by new online resources for delivering more meaningful and usable information to a full array of stakeholders. Recently garnered additional recognition from Achieve and Data Quality Campaign.

SLDS Examples from Other States

State: Arkansas **Organization(s): Arkansas Dept of Higher Education (ADHE), Arkansas Dept of Workforce Education (ADWE), Arkansas Research Center (ARC)** **Date of Snapshot: 2/1/2011**		
Context		
1. Has the state implemented or is it developing a P-20 data system or warehouse?	Yes	Distributed data system – plans to build a P-20 data warehouse. • Arkansas Research Center (ARC) at University of Central Arkansas serves as clearinghouse for linked state agency educational data. (See SHEEO Profile: http://sharepoint.dis.wa.gov/ofm/projects/ERDC/SLDS%20Projects%20by%20State/Arkansas/Profile%20of%20AR%20SLDS.pdf)
2. What are the main issues the P-20 system will enable the state to address?	→	• Primary focus: "Deliver data to the teacher." • Promoting use of SLDS data and creating a culture of data driven decision making (DDDM) throughout all levels of state education system and related government areas (e.g. workforce services) ○ ADHE has two partnering regional education cooperatives (Assessment and Accountability Comprehensive Center and Mid-Continent Comprehensive Center) with common goal of "build capacity among administrators, teachers and staff of cooperatives in Arkansas to increase effective use of school-based data for improved student learning" – resulted in Hive data visualization system/portal. • Teacher-student data link and tools delivered to teachers and school administrators to improve outcomes at school and district level. • Support Arkansas Education to Employment Tracking and Trends Initiative (AEETT). (See: Arkansas SLDS grant proposal 2009 - http://nces.ed.gov/programs/slds/pdf/Arkansas2009-ARRA.pdf).
3. Does the P-20 system have sufficient linked data available to address the issue(s) in 2.?	Yes	Access to P-20 system data is being provided for a wide range of purposes and stakeholders through at least four portals designed for major audiences and purposes— • "Hive" – Open and password-protected access to different views and levels of data. http://hive.arkansas.gov/ (See: Webinar by Hive designer - • "QuickLooks" – interactive tool for viewing district, school and county-level information. http://quicklooks.adearc.com/ • Department of Education Data Center portal—http://adedata.arkansas.gov. • ARC website – under development – intended to become central resource for research requests, data use management http://www.adearc.com/arc/index.html
4. Are the data being used to address the key issues?	Yes	Arkansas P20 resources are being used extensively for reporting, quality improvement and research, with more outreach and tools development underway. • As one example (reporting), by 2008 Arkansas had all four SLDS elements in place required by federal policy by 2011 for calculating graduation rates. (See: www.all4ed.org/files/Arkansas_grp.pdf)
5. Who is obtaining data from the P-20 system?	→	All education stakeholders, through portals and processes described in 3. • As example, see ADE data center portal for extensive list of types of information available depending on role of

		requestor and purpose. ○ Password-protected access and customized data views available for teachers, guidance counselors, principals and superintendents.
6. What kind of P-20 data are being used? - including degree to which personally identifiable, aggregate, linked, etc.	→	Aggregate and unit-record-level data accessible depending on identity, role and purpose of persons requesting access. • For example, ADE data center portal enables teachers and other school staff and administrators access to student-level data relevant for their responsibilities while preventing access to data beyond each individual's scope/purpose. • All portals provide interactive/customizable views of aggregated data to the extent allowed by state and federal privacy/confidentiality laws.
7. Who is qualified to request access to data?	→	All education stakeholders, with data available dependent upon identity, role and purpose of request.
8. Is there a formal policy or procedure for data sharing? Any governing legislation, policy, memorandum of understanding, etc.?		State has multiple formal policies and procedures within its distributed system – a few examples - • Multi-agency MOU between Dept of Higher Education, DOE and Department of Workforce Education – sets out process for data sharing, proposing and approving data use in research, general terms of use. (See: http://www.dataqualitycampaign.org/resources/details/710) • K-12 data security policy – sets out definitions and processes for handling "sensitive data," requires training for involved staff, requirements for secure transfer of data, etc. (See: http://arkedu.state.ar.us/commemos/attachments/ADE_K-12_IT_Security_Policy.pdf) • In 2009, Arkansas passed legislation creating a 13-member school leadership coordinating council and requiring it to devise a system for gathering education data and input on data use from a variety of stakeholders. (See: http://www.wallacefoundation.org/pages/main-report-strong-leaders-strong-schools.aspx#data-systems)
Details of Data Sharing Process for Research		
1. What is the process for requesting P-20 data?	Data may be requested through more than one P20 partner, but state is working toward establishing ARC at U of Central Arkansas as central clearinghouse. • ARC portal has a link to "data request" that produced an online form – but process is not fully automated yet. (See: http://www.adearc.com/arc/datarequest.html)	
2. How is the request reviewed and approved or denied? What criteria are applied?	• The online data request form indicates the ARC data team does an initial review of request to determine if requested data exist and whether they can be provided "within confines of state and federal laws."	
3. What restrictions and requirements are placed on the use and reporting of data? How communicated and enforced?	Formal policy/process is currently being developed. • According to email contact with ARC (2/2011), details of formal process for P20 sharing through ARC "clearinghouse" are under development.	
4. What protections are in place for protecting student confidentiality?	• Password protected, automated portals provide access to datasets, reports, customizable data views –aggregated or unit-level—depending upon identity, role and purpose of data requestor/user. • In the past, ADE policy was no unit-level record data shared with external researchers – currently, through ARC, requestor requests are being allowed but are being processed on case-by-case basis by research team.	
5. How much time and cost	Depends upon details of request.	

are involved in requesting and receiving data?	• As of 2/1/2011, no standard, written policy available – time and cost based on case-by-case analysis. • Standard policy and process is under development.

SLDS Examples from Other States

State: Florida **Organization(s): Florida Department of Education (FDE)** **Date of Snapshot: 1/13/2011**		
Context		
1. Has the state implemented or is it developing a P-20 data system or warehouse?	Yes	The Florida K-20 Education Data Warehouse (EDW): • Integrates existing, transformed data extracted from multiple sources that are available at the state level. • Provides a single repository of data concerning students served in the K-20 public education system as well as educational facilities, curriculum and staff involved in instructional activities. (See EDW Factsheet, http://edwapp.doe.state.fl.us/EDW_Facts.htm.)
2. What is the main reason for the state to invest in P-20 resources?	→	To answer the following questions: 1. What is the public receiving in return for funds it invests in education? 2. How effectively is Florida's K-20 education system educating its students? 3. How effectively are the major delivery sectors promoting student achievement? 4. How are individual schools and postsecondary education institutions performing their responsibility to educate their students as measured by how students are performing and how much they are learning? (See 2010 Florida Statute 1008.31, http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=1000-1099/1008/Sections/1008.31.html.)
3. Does the P-20 system have sufficient linked data available to address the issue(s) in 2?	Yes	Linked or linkable data resources, going back as far as 1995 • Public pre-K through graduate school • Student-level data for public schools, community colleges, career and technical education, adult education, and state university system • Staff, facilities, finance, financial aid • Post-school employment and non-education system program data • As of 10/2010, FL announced it is further developing its early learning data system with plans to link into P20 SLDS. (See SREB/DQC Sellers Presentation, 2008, http://www.dataqualitycampaign.org/files/Sellers_Presentation_Nov_19_2008.pdf)
4. Are the data being used to address the key issues?	Yes	Being used for research, evaluation, assessment and improvement in many issue areas, including: • Teacher preparation and development best practices, compensation, results of retention policies • CHOICE Option evaluations • Return on investment • Evaluating key transitions for all students • Acceleration mechanisms • SAT, ACT, testing anomalies • Results of retention policies (See "Using State Longitudinal Education Data to Drive System Performance," Pfeiffer presentation, 2010, http://www.nga.org/Files/pdf/1004EDUINSTITUTEPFEIFFER.PDF.)

5. Who is obtaining data from the P-20 system?	→	• Local and state education administrators, evaluators, practitioners • Education consumers and other public stakeholders • State and national researchers, centers and networks including CALDER, Community College Research Center, etc. (See Pfeiffer presentation.)
6. What kind of P-20 data are being used? - including degree to which personally identifiable, aggregate, linked etc data included.	→	• Wide range of K-12, Pre-K and postsecondary student, staff, curriculum and institution data • Workforce and other social services and health data relevant to outcomes of education • Unit-record-level and aggregated data available as appropriate to role of researcher and type of project (See map of all data types being linked - Pfeiffer presentation)
7. Who is qualified to request data (criteria)?	→	• "Including, but not limited to, administrators, educators, parents, students, state leadership, and professional organizations" • Criteria for accessing data depend on who the requestor is, what data are requested, and intended use.
8. Is there a formal policy or procedure for data sharing? Describe any governing legislation, policy, memorandum of understanding, etc.	Yes	• 2010 Florida statute regarding P-20 system and data: 1008.31 • Data sharing through Memoranda of Understanding (MOUs) between state agencies including workforce, corrections, children and family services, and federal organizations including defense and National Student Clearinghouse. (See Sellers presentation.)
Details of Data Sharing Process for Research		
1. What is the process for requesting P-20 data?	For requesting use of unit record data: • Complete and submit unit record request packet (project information, description, timeline, statement of benefit, data element crosswalk) (See http://www.fldoehub.org/Research/Documents/Unit%20Record%20Data%20Request%20Packet%20Instructions.doc.)	
2. How is the request reviewed and approved or denied? What criteria are applied?	• Request forms are logged into tracking system • Reviewed by staff, possibly reviewed by committee, depending on details of request. • General priorities for request processing are: legislative and gubernatorial requests completed first, internal requests second, and external requests third.	
3. What restrictions and requirements are placed on the use and reporting of data? How communicated and enforced?	• If request for data is approved, researcher signs a Security and Access Agreement that sets out conditions of data use for particular project. • Agreement can be revoked for non-compliance with any terms. • Full description of formal policy and procedures, and forms for applying for and using data, are available through data hub: http://www.fldoehub.org/Research/Pages/default.aspx	
4. What protections are in place for protecting student confidentiality?	• In general, Florida education data sources have embedded SSNs that are matched within secure data environment and replaced in released datasets by a unique identifier. • Florida policy, data access portal, processes and training refer to privacy and confidentiality regulations, and determine what level of data are available to various stakeholders. • Confidentiality and data security requirements are included in Security and Access Agreement signed by researchers, included in the unit record data request packet – attached.	

5. How much time and cost are involved in requesting and receiving data?	• Minimum of 3-4 weeks for approval decision, and up to 5-6 months to receive data requested. • Time and cost "vary greatly by individual proposal depending upon data permissions required, datasets requested [whether standard or requiring customization], and the number of proposals currently approved."

SLDS Examples from Other States

State: Kansas **Organization(s): Kansas Board of Regents-KBOR (post-secondary), Kansas Dept of Education-KDE (K-12), Kansas Education Data Users Consortium-KEDUC (includes KBOR, KDE, University of Kansas-KU and Kansas State University-KSU)** **Date of Snapshot: 1/13/2011**		
Context		
1. Has the state implemented or is it developing a P-20 data system or warehouse?	Yes	In the process of building—immediate goals: • Expand K-12 and post-secondary systems to link across the P-20 education pipeline and across state agencies. • Ensure education data can be accessed, analyzed and used by stakeholders to promote continuous improvement. • Build capacity of educators to use longitudinal data for effective practice, to inform instructional decisions and evaluate effects on student learning. (See NCES SLDS grant summary, http://nces.ed.gov/Programs/SLDS/state.asp?stateabbr=KS)
2. What are the main issues the P-20 system will enable the state to address?	→	• Better student outcomes (behavior, achievement, persistence in education) from PreK through postsecondary, especially for challenged groups (disabilities, limited English proficiency) and schools (low-performing). • Improve educator competency and professional satisfaction. • Evaluate and increase effective use of education data systems by teachers and principals for improvement. (See NCES SLDS grant summary)
3. Does the P-20 system have sufficient linked data available to address the issue(s) in 2.?	No	K-12 and post-secondary data have not yet been linked; as of Jan 2011, reorganization of governance and data sharing structure is taking place. (See P-20 Committee's report, Dec 2010: http://www.ksde.org/LinkClick.aspx?fileticket=zmSQS0DUgC0%3d&tabid=2880&mid=10681)
4. Are the data being used to address the key issues?	No	Not yet—still in planning and early implementation stages as reported in P-20 Committee's report: • "KSDE and KBOR have signed a Memorandum of Understanding allowing mutual access to student data that, with proper security controls, makes it possible to track the progress of students as they move through the educational system." • Integration with Departments of Labor and Commerce data will follow the linking of K-12 and post-secondary data and creation of data warehouse.
5. Who is obtaining data from the P-20 system?	→	No one; P-20 warehouse not completed. Some research going on using K-12 data. • In Jan 2011, KBOR and KEDUC reported researchers continue to go directly to K-12 and university data sources instead of through KEDUC, in spite of incentives to steer them to centralized resource—probably due to lack of awareness of potential P-20 data uses combined with very limited P-20 data available so far. • P-20 steering committee is reorganizing to realign outreach activities with current, early stage of data warehouse, and implementing a one-on-one outreach strategy with P-20 partner organizations' researchers only.
6. What kind of P-20 data are being used? - including degree to which personally identifiable, aggregate, linked, etc., are included.	→	None yet—as of Dec 2010, P-20 warehouse still being developed. • K-12 "restricted-use data" have been shared for some time by KDE on case-by-case basis and linked by researchers, and more recently made available through KEDUC formalized process, for evaluation, reporting and accountability. ○ KS definition of "restricted-use data" is "all data containing personally-identifiable information collected by or on behalf of KSDE and/or KBOR that are provided to the Researcher and all information derived from those data, and all data resulting from merges, matches, or other uses of the data provided by KSDE/KBOR

		with other data." • Details of KDE K-12 data and sharing processes can be seen through KDE portal, separate from KEDUC (See http://www.ksde.org/Default.aspx?tabid=83)
7. Who is qualified to request access to data?	→	• Once P-20 data become available, according to KEDUC website: A principal investigator (PI) who wants to use state education data to conduct research related to stakeholder-identified research priorities. • In Jan 2011, KBOR indicated data sharing access and support will be scaled back to only researchers from P-20 partners—not other state agencies, smaller education institutions, external researchers etc.
8. Is there a formal policy or procedure for data sharing? Any governing legislation, policy, memorandum of understanding, etc.?	Yes	• In June 2009, in response to urging from KBOR and KSDE, Governor signed Executive Order 09-03 allowing the two agencies to cross-evaluate programs and share student data. (See http://www.ksde.org/LinkClick.aspx?fileticket=zmSQS0DUgC0%3d&tabid=2880&mid=10681 and http://kasa-ks.org/council/Council%20meeting%20materials%2010-21/MOU.pdf) • Formal policy and process have been developed and is implemented through KEDUC website, but as of Dec 2010 have not yet been used. (See http://www.keduc.info/proposal/background)
Details of Data Sharing Process for Research		
1. What is the process for requesting P-20 data?	Currently in use for available data (K-12) and will also be used for P-20 data when data warehouse is ready: • Through KEDUC website, researchers initiate a request, reviewed by KEDUC staff. • "Principal investigator advocate" (PIA) determines type of data needed, where it can be obtained and, if restricted-use data are required for the proposed project, assists the PI in completing more detailed research proposal forms and data requests as needed, then assists with submitting to Review Committee.	
2. How is the request reviewed and approved or denied? What criteria are applied?	• Website provides password-protected access for Review Committee members to collaboratively review/approve request. • KEDUC Research Committee reviews and makes approval decision based on pre-established criteria and rating system. (See attached, or at http://www.keduc.info/images/keduc-criteria-chart-large.jpg) • After approval, PIA assists PI in finalizing research agreement forms, any further necessary data requests or agreements, and tracking of agreement performance.	
3. What restrictions and requirements are placed on the use and reporting of data? How communicated and enforced?	• Restrictions are communicated through a Research Project Confidentiality Agreement. Agreement may be revoked if any conditions are violated. For details see attached, or http://www.keduc.info/proposal/sign • PIs with approved proposals must provide KEDUC with research results, including an abstract, when the research is completed, and PIs may be asked to present results at a state education workshop/conference	
4. What protections are in place for protecting student confidentiality?	• Kansas has developed and is using unique personal identifiers instead of SSN, in preparation for P-20 linking. • Other specific protections for student confidentiality and data security are detailed in the Research Project Confidentiality Data Use Agreement; for full details, see attached, or on KEDUC website, http://www.keduc.info/proposal/sign	
5. How much time and cost are involved in requesting and receiving data?	• To provide incentive for researchers to use KEDUC, KDE's and KBOR's $60/hour fee for preparing datasets is waived if researcher applies through KEDUC. • Time estimate for reviewing proposals and delivering data varies depending on project priority, amount of data preparation required, and staff workload.	

	• Proposals submitted through KEDUC that address state's target issues get priority.

KEDUC CRITERIA FOR ACCEPTING RESEARCH TO SUPPORT

Factor	Poor (1)	Fair (2)	Good (3)	Excellent (4)	Score (Rating * factor weight)
Research productivity of principal investigator **Weight: 5**					
Principal investigator who is an employee of a Kansas accredited education agency	NA	NA	NA	NA	Score: No=0 Yes=20
Alignment of research questions to stakeholder-identified research priorities **Weight: 10**					
Clarity of research objectives **Weight: 10**					
Alignment of research objectives with research design **Weight: 10**					
Opportunity of research to add to knowledge base/impact education practice or policy **Weight: 10**					
Appropriate use of state longitudinal data **Weight: 10**					
Likelihood of successful completion within the stated timeline **Weight: 5**					

RESEARCH PROJECT CONFIDENTIALITY AGREEMENT
(From: Kansas Education Data Users Consortium)

WHEREAS, the Kansas State Department of Education (KSDE) and/or the Kansas Board of Regents (KBOR) have collected certain data that contain confidential personally-identifiable information, and KSDE and KBOR require this confidentiality to be protected; and

WHEREAS, the Kansas State Department of Education and Kansas Board of Regents are willing to make these data available for research and analysis purposes to improve instruction in public elementary and secondary schools and postsecondary schools, but only if the data are used and protected in accordance with the terms and conditions stated in this Agreement.

NOW, THEREFORE, it is hereby agreed between

(Typed name and address of Research Organization), hereinafter referred to as the "Researcher," and KSDE and/or KBOR that:

I. INFORMATION SUBJECT TO THIS AGREEMENT

A. All data containing personally-identifiable information collected by or on behalf of KSDE and/or KBOR that are provided to the Researcher and all information derived from those data, and all data resulting from merges, matches, or other uses of the data provided by KSDE/KBOR with other data, are subject to this Agreement and are referred to herein as "restricted-use data." The restricted-use data under this Agreement may be provided in various forms included but not limited to written or printed documents, computer tapes, diskettes, CD-ROMs, hard copy, or encrypted files.

The Researcher may use the restricted-use data only for the purposes stated in the Research Proposal Application, which is attached hereto and made a part of this Agreement (marked as Attachment 1), and is subject to the limitations imposed under the provisions of this Agreement.

II. INDIVIDUALS WHO MAY HAVE ACCESS TO RESTRICTED-USE DATA

Researcher agrees to limit and restrict access to the restricted-use data to the following three categories of individuals:

The Project Leaders in charge of the day-to-day operations of the research and who are the research liaisons with the Principal Investigator Advocate (PIA).

The Professional/Technical staff in charge of the research under this Agreement.

Support staff including secretaries, typists, computer technicians, etc., but these individuals shall be allowed access to the restricted-use data only to the extent necessary to support the research.

III. LIMITATIONS ON DISCLOSURE

A. Researcher shall not use or disclose the restricted-use data for any purpose not expressly stated in the Research Proposal Application approved by KEDUC, unless the Researcher has obtained advance written approval from the PIA.

B. Researcher may publish the results, analysis, or other information developed as a result of any research based on the restricted-use data made available under this Agreement only in summary or aggregate form, ensuring that the identities of individuals included in the restricted-use data are not revealed.

IV. ADMINISTRATIVE REQUIREMENTS

A. The research conducted under this Agreement shall be limited to, and consistent with, the purposes stated in the Research Proposal Application.

B. Notice and training on confidentiality and nondisclosure.

1. Researcher shall notify and train each of its employees who will have access to the restricted-use data of the strict confidentiality of such data, and shall require each of those employees to execute an Acknowledgement of Confidentiality Requirements.

2. Researcher shall maintain each executed Acknowledgement of Confidentiality Requirements at its facility and shall allow inspection of the same by the PIA upon request.

3. Researcher shall promptly notify the PIA in writing when the access to the restricted-use data by any individual is terminated, giving the date of the termination.

C. Publications made available to KEDUC.

1. Copies of each proposed publication or document containing or based upon the restricted-use data shall be provided to the PIA before the publication or document is finalized. The PIA will share the proposed publication with the appropriate agency or agencies (KSDE and/or KBOR) so that a restricted-use data security review can be performed. The PIA shall advise the Researcher when publication is authorized.

2. Researcher shall provide KEDUC a copy of each publication based on the restricted-use data made available with KEDUC assistance.

D. Researcher shall notify the PIA immediately in writing upon receipt of any request or demand for disclosure of restricted-use data.

E. Researcher shall notify the PIA immediately in writing upon discovering any breach, or suspected breach, of security, or of any disclosure of restricted-use data to an unauthorized party or agency.

V. SECURITY REQUIREMENTS

A. Maintenance of, and access to, the restricted-use data.

1. Researcher shall retain the original version of the restricted-use data at a single location and shall not make a copy or extract of the restricted-use data available to anyone except individuals specified in paragraph II.

2. Researcher shall maintain the restricted-use data (whether maintained on a mainframe facility, central server, personal computer, or in print or other medium materials) in an area that has access limited to authorized personnel only. Researcher shall not permit removal of any restricted-use data from the limited access area.

3. Researcher shall ensure that access to the restricted-use data maintained in computer files or databases is controlled by password protection. Researcher shall maintain all printouts, diskettes, or other physical products containing restricted-use data in locked cabinets, file drawers, or other secure locations when not in use.

4. Researcher shall ensure that all printouts, tabulations, and reports are edited for any possible disclosure of restricted-use data.

5. Researcher shall establish procedures to ensure that the restricted-use data cannot be extracted from a computer file or database by unauthorized individuals.

B. Retention of restricted-use data.

1. Researcher shall destroy the restricted-use data, including all copies, when the research that is the subject of this Agreement has been completed or this Agreement terminates, whichever occurs first.

VI. TERMINATION OF THIS AGREEMENT

This Agreement shall terminate twelve months from the date it is signed by the researcher and PIA. The Agreement, however, may be extended by written agreement of the parties.

Any violation of the terms and conditions of this Agreement may result in the immediate revocation of this Agreement. The PIA may initiate revocation of this Agreement by written notice to Researcher indicating the factual basis and grounds of revocation. Upon receipt of the written notice of revocation, the Researcher shall immediately cease all research activity related to the Agreement until the issue is resolved. The Researcher will have 3 business days to submit a written Response to the PIA indicating why this Agreement should not be revoked. The appropriate agency's review board (KSDE's or KBOR's) shall decide whether to revoke this Agreement based on all the information available to it. The PIA shall provide written notice of the agency's decision to the Researcher within 10 business days after receipt of the Response. These timeframes may extend for good cause.

Download Signature Page

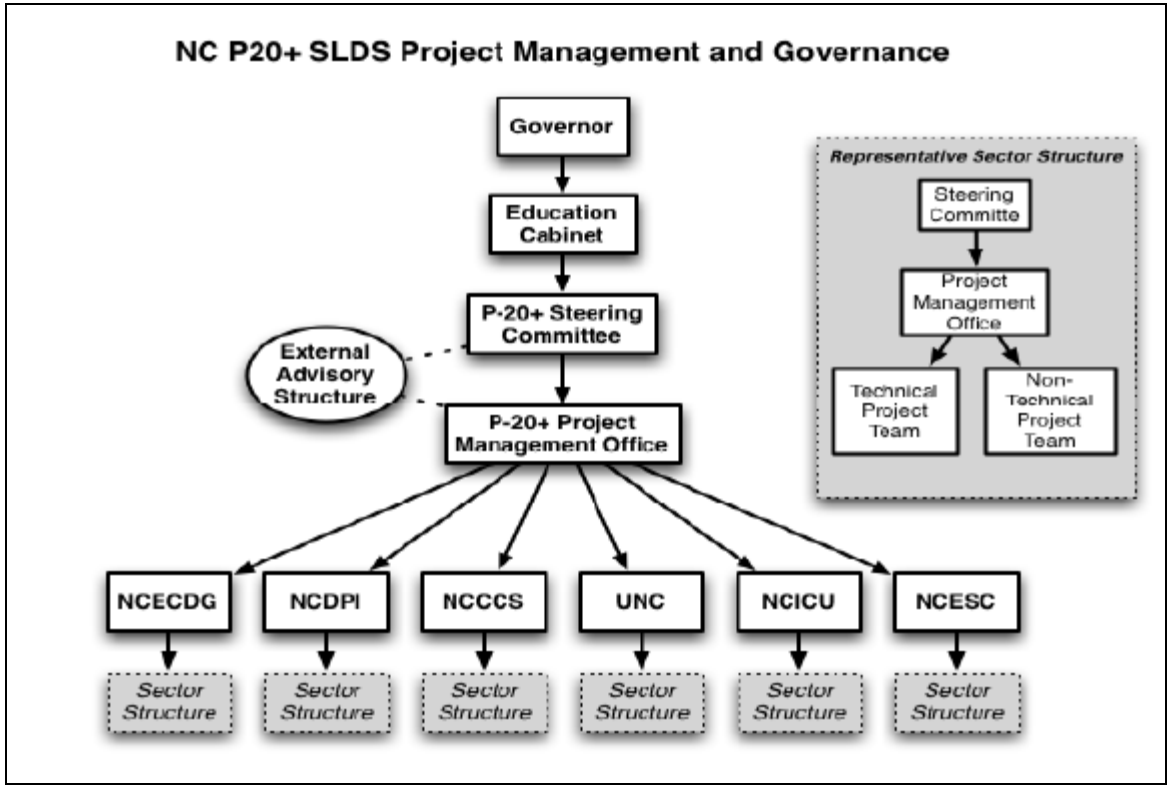
SLDS Examples from Other States

State: North Carolina **Organization(s): NC Dept of Public Instruction (NCDPI), University of NC system (UNC), NC Early Childhood Data Group (NCECDG), NC Community College System (NCCCS), NC Independent Colleges and Universities (NCICU), NC Employment Security Commission (NCESC)** **Date of Snapshot: 1/13/2011**		
Context		
1. Has the state implemented or is it developing a P-20 data system or warehouse?	Yes	In the process of building and implementing a P-20+ distributed data system by linking existing source data systems from numerous partners (see above), plus P-20 governance structure including numerous stakeholders including at least one education research data center (North Carolina Education Research Data Center, or NCERDC, at Duke University)
2. What are the main issues the P-20 system will enable the state to address?	→	Enable "NC leaders at all points along the NC education-workforce continuum access to a broader view of the State's educational needs…as NC strives to find the right formula(s) for ensuring that our State's spectrum of education services can facilitate every student's achievement of college- and/or career-readiness." (See 2009 SLDS grant application: http://nces.ed.gov/programs/slds/pdf/NorthCarolina2009-ARRA.pdf)
3. Does the P-20 system have sufficient linked data available to address the issue(s) in 2.?	Yes and No	YES: Among separate state partners, there are considerable education data resources that already being linked and used for evaluation and research—but NO: Not yet through centralized P-20 system. • As of Dec 2010, most of these resources and longitudinal data sharing and linking activity still take place outside the planned, coordinated P-20 data warehouse and governance structure. • NC's P-20+ project is primarily focused on need to establish centralized governance and data integration to streamline, improve efficiency, and realize full potential of the state's collective data resources.
4. Are P-20 data being used to address the key issues?	Yes	For years, North Carolina has provided in-state and out-of-state researchers with linked, longitudinal data resources, including through a partnership between NCDPI and NCERDC. • P-20+ will incorporate, coordinate and leverage existing productive partnerships and data repositories, to enable more data-driven decision making.
5. Who is obtaining data from the P-20 system?	→	• New centralized P-20+ system: None, implementation of P-20+ governance system and warehouse not yet complete. • From various, uncentralized partners of what will be the P-20+ system: An array of public, state and national agency, and local, regional and national research stakeholders. For example - ○ From NC WISE, K-12 system that completed rollout to all 115 NC LEAs and 98 charter schools in 2009 – providing linked, longitudinal information to individual schools, school districts, universities and colleges, and NCDPI. ○ From CEDARS, NC DPI's new PreK-13 SLDS – has a student and staff identification system that will be used for full P-20+ linking, centralized data repository, reporting and analysis tools providing access to state, local and federal policymakers and service providers, and information sharing among DPI staff, school principals, LEA administrators, and the public. ○ From NCERDC linked data warehouse—state researchers from North Carolina and other states, national-level research centers and networks (see more details below).
6. What kind of P-20 data are being used? - including degree to which	→	• Limited amount of linking and sharing of P-20+ partner/sector data already taking place through NCDPI and NCERDC. • When P-20+ project completes linking of significant data sources of partners, it will encompass a full array of unit-

personally identifiable, aggregate, linked, etc., are included.		record-level, aggregated and linked data covering PreK through postsecondary and beyond to workforce. • P-20 project status identifies a few remaining data gaps mostly related to interconnectability and quality issues.
7. Who is qualified to request access to data?	→	Full array of education stakeholders, who currently go through several different channels to request and access data and information. Refer to description of existing data portals in 5 above.
8. Is there a formal policy or procedure for data sharing? Any governing legislation, policy, memorandum of understanding, etc.?	Yes and No	• YES: For the most part, the major partners' centralized data resources have formal policies and procedures for accessing education data and information. • NO: P-20+ governance and sharing processes are still under development—expected to incorporate and expand on work already developed by major partners, including NCDPI and NCERDC.
Details of Data Sharing Process for Research		
1. What is the process for requesting P-20 data?		Focusing on process of North Carolina Education Research Data Center (NCERDC): • Researcher goes to NCERDC web portal to obtain information about data available, criteria for using data, process for requesting data, and forms to apply for data access. (for details, see http://www.childandfamilypolicy.duke.edu/project_detail.php?id=35)
2. How is the request reviewed and approved or denied? What criteria are applied?		• Data can be released to an institution of higher education, a non-profit research institution, or a government agency with established protocols for an Institutional Review Board (IRB), Human Subjects Review Committee, or equivalent body. • Researcher must have primary affiliation with an eligible institution or be a currently enrolled doctoral student. NCERDC's list of eligible institutions and researchers continues to grow, and now includes national-level research centers and networks as well as in-state. • Prioritization of requests: ○ First priority goes to State Board of Education-related research. ○ Second priority to proposals that fit NCERDC's goals of supporting research useful for education policy (detailed description on their website). ○ Other priority considerations include North Carolina-based, likelihood of benefits to NCERDC, and amount of staff time required by project. • Once request reviewed by NCERDC review committee and approved, process for obtaining and using data includes— ○ Complete and sign a data use agreement (http://www.childandfamilypolicy.duke.edu/pdfs/projects/NCERDC_DataUseApplication.doc) ○ Sign Investigator confidentiality agreement and ensure all research assistants sign and abide by conditions of agreement. ○ Obtain their institution's IRB approval of project (expedited or full review). ○ Complete Data Request Form (included online Data Use Application packet)
3. What restrictions and requirements are placed on the use and reporting of data? How		• Follow plan approved by NCERDC and the researcher's IRB. • Provide NCERDC with any resulting "paper"—including conference presentations, accepted publications, press releases. • Communicated and enforced through signed data use agreement, which may be revoked for failure to comply with terms.

communicated and enforced?	
4. What protections are in place for protecting student confidentiality?	• Approval of project and delivery of requested data is contingent upon NCERDC's and researcher's ability to construct adequate datasets while complying with FERPA and all other relevant privacy and confidentiality regulations. • Processes for protecting access to data are detailed in data use agreement specific to the approved project. • IRB must review and approve plan for data use and confidentiality protection, as well as NCERDC.
5. How much time and cost are involved in requesting and receiving data?	• No time estimate for processing request or delivering data is provided; instead, NCERDC publicizes its prioritization rules on website. • NCERDC charges a standard fee of $1,800, "equivalent to two days of NCERDC services." Fee covers limited telephone and email consultation with investigator or research staff about origins, structure and general content of data files sent. Fee may be waived for doctoral candidates and may be negotiable for some faculty projects. Additional fees may be charged for customized datasets, and NCERDC will provide researcher with estimate of time and associated fees.

Proposed P-20+ central governance structure for distributed (rather than centralized) education data system



SLDS Examples from Other States

State: Pennsylvania **Organization(s): Pennsylvania Department of Education (PDE); Pennsylvania Office of Childhood Development and Early Learning (OCDEL)** **Date of Snapshot: 1/28/2011**		
Context		
1. Has the state implemented or is it developing a P-20 data system or warehouse?	YES	Pennsylvania Information Management System (PIMS) is PDE's data collection and enterprise-wide longitudinal data warehouse and reporting system. • System currently has longitudinal student and teacher IDs with student/teacher match; test, enrollment, participation and program data; data on students not taking tests; high school dropout and graduation data; K12 and postsecondary data links; data audit system. • In a separate system (Early Learning Network or ELN), state has a nationally recognized integrated database for PreK that is linkable by state's unique student ID. • Plans for 2010-2013: Expansion of PreK and K12 data collection; reporting for principals (Cognos cubes); expand postsecondary implementation; link early learning data network (ELN) to PIMS; link PIMS to National Student Clearinghouse data; begin implementation of PDE data access and use policy; retire legacy data systems. (See: PIMS portal - http://www.portal.state.pa.us/portal/server.pt/community/pims_-_postsecondary/8960; and ELN portal - http://www.portal.state.pa.us/portal/server.pt/community/early_childhood/8705/early_childhood_care_and_education_research/522247; and PA SLDS grant proposal, http://nces.ed.gov/programs/slds/pdf/Pennsylvania2009-ARRA.pdf)
2. What are the main issues the P-20 system will enable the state to address?	→	• Meeting current state and federal reporting requirements • Improving education decision-making through the use of high quality data and decision support tools • Providing longitudinal tracking of particular individual and subgroup education progress over time and across LEAs • Reporting timely and accurate education data through standardized and ad hoc reporting capabilities. (See PIMS portal)

3. Does the P-20 system have sufficient linked data available to begin addressing the issue(s) in 2.?	YES	• As of Dec 2009, PIMS had expanded its K12 system to link data from all 14 Pennsylvania public higher education institutions.
		  ○ PDE functions as custodian and policy administrator of all data resources in the linked system.
		• PA's Early Learning Network (ELN) has received several citations, including NGA Best Practices, as example for integrated early childhood data system—it is being used for reporting and decision-making for some EL issues.
		  ○ PA currently is working on resolving data access issues across the early childhood and K12 data systems, including privacy issues, to enable secure P20 linking and data sharing.
		(See PIMS portal, and Building Ready States: Governor's Guide to Early Childhood State Systems http://www.nga.org/portal/site/nga/menuitem.9123e83a1f6786440ddcbeeb501010a0/?vgnextoid=9d00bc9a03cdb210VgnVCM1000005e00100aRCRD)

4. Are the data being used to address the key issues?	YES	PDE is using linked data for research on key policy, quality and compliance issues, by state agency researchers and those third-party researchers who have been authorized to conduct research on behalf of the state agencies.
5. Who is obtaining data	→	State agency and authorized third-party researchers who are using data for the limited purposes set out in 2 above.

from the P-20 system?		
6. What kind of P-20 data are being used? - including degree to which personally identifiable, aggregate, linked, etc., are included.	→	Full array of student, teacher, school data are available for state education agency use, through PIMS and ELN. • Unit-level data and records are available to PA education system stakeholders depending on the identity and role of the requestor logging in to the ID-secured portal (PASecureID), or requesting data on case-by-case basis through PDE administration. • Early childhood data network (ELN) enables "access to data and reports based on a person's level of oversight of children and operations" – see description in NCSL case study on PA early childhood data system.
7. Who is qualified to request access to data?	→	All state education staff, students and parents, other stakeholders including public—if complying with defined and approved reasons for accessing and using data. • Type of data and whether unit-level or aggregated depends on identity, role and purpose for accessing data. • PDE's general policy for access by outside researchers is "the Department may not redisclose personally identifiable information to a third party researcher unless the researcher is acting as an 'authorized representative' of the Department, acting under the control of the Department as an employee, appointed official or contractor who is providing services that the Department would otherwise provide for itself." o PDE must have authorized the study and it must be conducted for or on behalf of the Department. o "That fact that an outside entity, on its own initiative, conducts a study which may benefit an educational agency or institution does not transform the study into one done 'for or on behalf of' the Department." (See PDE Student Data Access and Use Policy – http://www.google.com/url?sa=t&source=web&cd=1&ved=0CBsQFjAA&url=http%3A%2F%2Fwww.dgs.state.pa.us%2Fportal%2Fserver.pt%2Fgateway%2FPTARGS_0_2_637118_0_0_18%2FPIMS_Data_Access_Policy.pdf&ei=0mJDTc2XO4a-sAOGoJnzCg&usg=AFQjCNGUNgZkyVuF9jBORDH8SUE6yM3QjQ)
8. Is there a formal policy or procedure for data sharing? Any governing legislation, policy, memorandum of understanding, etc.?	YES	• Access to aggregated and individual-level data for many types of data users through secure-access portals. • Inter-agency agreements are used to set out and enforce terms of data sharing among state agencies for the purposes shown in 2 above. (See: PDE Student Data Access and Use Policy)
Details of Data Sharing Process for Research		
1. What is the process for requesting P-20 data (beyond what is available through portals)?		Contact PDE or early childhood department staff and submit request per instructions. • PDE indicates in their data access and use policy, "The Department will work with researchers with the goal that they receive the most meaningful data possible without the disclosure of information that would make any student's identity easily traceable." (See: PDE Student Data Access and Use Policy)
2. How is the request reviewed and approved or denied? What criteria are applied?		• Requests are considered on a case-by-case basis by PDE administrator and staff to determine whether they are in compliance with state and federal laws and regulations.
3. What restrictions and		• Any release of student data to researchers outside PDE is considered a loan of data, i.e., the recipients do not have

	requirements are placed on the use and reporting of data? How communicated and enforced?	ownership of the data. • Researchers are required to supply a copy of any analysis or reports created with the data. • Researchers are required to destroy the data once the research is completed.
4.	What protections are in place for protecting student confidentiality?	• Assignment of a unique identifier (PASecureID) – 10-digit number maintained for every Pennsylvania student such that only one student is ever assigned a particular number; the number is always associated with that student throughout the educational career or until leaving the state; a student is only ever assigned one number, preventing duplication. • Data security – Technical measures put in place by the State to make sure records "are not lost, stolen, vandalized, illegally accessed or otherwise rendered useless" including secure firewalls, secure socket layers, audit trails and physical security such as restricted room access; regular state and federal auditing. • Restricted access to student-level data, according to identity, role and purpose for data access (see details in Student Data Use and Access Policy). • Statistical security – As a general policy, individual student data will be aggregated to comply with required state and federal reporting. "When there is a risk of statistical disclosure such as in the case of narrowly defined or small populations, the Secretary of Education will enforce statistical cutoff procedures using a minimum confidentiality *n* of 10" or otherwise as necessary to maintain confidentiality. (See PDE Student Data Access and Use Policy)
5.	How much time and cost are involved in requesting and receiving data?	• Time to approval and delivery of data depends on alignment with PDE research and policy priorities, and staff workload. • PDE "reserves the right" to charge "a reasonable fee" for the use of data by researchers to help offset the state's costs of collecting and storing the data. (See PDE Student Data Access and Use Policy)

Note: In October 2010, Pennsylvania received nationwide publicity for a major data breach of state public welfare and medical records. Even though the breach did not occur in any state education data system, a state legislator reacted to try to halt development of the education systems.

See:

Medical data breach said to be major
http://www.philly.com/inquirer/business/20101021_Medical-data_breach_said_to_be_major

Pa., suspend risky system that collects student data (editorial by Jeffrey Piccola, Republican, 15th senate district)
http://www.philly.com/inquirer/opinion/20101114_Pa___suspend_risky_system_that_collects_student_data.html

SLDS Examples from Other States

State: Texas **Organization(s): Texas Education Agency (TEA), Texas Higher Education Coordinating Board (THECB), Texas Education Research Centers (ERC)** **Date of Snapshot: 2/14/2011**		
Context		
1. Has the state implemented or is it developing a P-20 data system or warehouse?	Yes	Texas was one of the first states to develop centralized education data system for compliance and reporting. • By 2008, TX's original education data system needed to be upgraded—supported by a grant from Dell Foundation, Texas contracted with IBM Global Business Services for an extensive, stakeholder-based needs assessment, followed by a multi-year implementation that is ongoing. (See: http://ritter.tea.state.tx.us/tea/IBM_TDCARSI_Recommendation.pdf
2. What are the main issues the P-20 system will enable the state to address?	→	• Use key data to better understand students' preparedness to contribute to the 21st century workforce. • Alleviate data collection burden on districts and improve data quality. • Build performance management culture that uses data to continuously improve performance.
3. Does the P-20 system have sufficient linked data available to address the issue(s) in 2.?	Yes	• The longitudinal data system website and portal points of the education research centers show currently linked data resources: http://www.texaseducationinfo.org/tpeir/ , http://utaustinerc.org/?sid=5&pid=51#s5 • The system is continuing to develop and release data resources, with plans to release in the near future: ○ student-to-teacher tracking information for PK-12 ○ SAT and ACT student performance data ○ National Student Clearinghouse information
4. Are the data being used to address the key issues?	Yes	• Examples of longitudinal data being used to answer policy and performance questions can be seen at: ○ Division of Accountability Research, TEA - http://www.tea.state.tx.us/acctres/dropcomp_index.html ○ Education Research Center at University of Texas – Dallas: http://www.utdallas.edu/research/tsp-erc/ ○ Education Research Center at University of Texas – Austin: http://www.utaustinerc.org/?sid=4&pid=41 ○ Education Research Center at Texas A&M - http://erc.cehd.tamu.edu/edprep.html
5. Who is obtaining data from the P-20 system?	→	• The TPEIR portal can be used by legislators, educators, parents and researchers to obtain data and reports. http://www.texaseducationinfo.org/tpeir/
6. What kind of P-20 data are being used? - including degree to which personally identifiable, aggregate, linked, etc., are included.	→	• Data are available at all levels from the system at levels from unit-level to aggregate as appropriate for to data requestor and purpose. • An online inventory of available data systems and data elements is available from Texas Education Research Center system: http://www.utaustinerc.org/inventory.php
7. Who is qualified to request access to data?	→	• A full array of education stakeholders are welcome to obtain information from the data system to address a wide variety of policy, performance and research questions. • Texas education research center (ERC) data may be used only for research projects specifically approved by the ERC Joint Advisory Board, and for investigative and analysis tasks upon direction of one or both JAB commissioners. (See: http://www.utaustinerc.org/files/Texas_ERC_Terms_and_Procedures.pdf)
8. Is there a formal policy or procedure for data	Yes	• In 2008, state legislature directed Texas Education Agency and Texas Higher Education Coordinating Board to establish three state education research centers (ERCs), and added rules for ERC operation to state administrative

sharing? Any governing legislation, policy, memorandum of understanding, etc.?		code. (See: http://www.capitol.state.tx.us/tlodocs/793/billtext/html/HB00001F.htm) • Each ERC executes a three-party interagency cooperative contract with TEA and THECB embodying legal framework and administrative requirements for data sharing and security. • THECB is lead agency for maintaining data stores for each ERC, depositing data from TEA and other state education agencies, and preparing data for use by variety of stakeholders. (See: Texas FERPA Story, http://www.urban.org/uploadedpdf/1001267_texasferpastory.pdf)
Details of Data Sharing Process for Research		
1. What is the process for requesting P-20 data?		• Researchers can request data through several portal points provided by the three Texas education research centers. • Researchers submit written requests for data review and approval, either through email or printed letter.
2. How is the request reviewed and approved or denied? What criteria are applied?		According to process publicized on the Education Research Center portal: • All requests for review and release of data are logged in ERC system. • All research results and products must be reviewed by the Director or Associate Director of the Texas ERC before release. (See: http://www.utaustinerc.org/?sid=5&pid=53)
3. What restrictions and requirements are placed on the use and reporting of data? How communicated and enforced?		• To obtain data online through ERC, researchers must apply for and obtain a University of Texas electronic identity (UT EID). Researchers without a UT EID must obtain assistance from an ERC staff member during normal business hours. • The restrictions and requirements are communicated through the ERC website and through written agreements specific to projects. • The terms are enforceable through documented agreements.
4. What protections are in place for protecting student confidentiality?		• Researchers must access the ERC data warehouse through approved, secure client workstations, either through use of a pre-obtained UT EID or through an ERC staff member. • Identifying data elements such as student names or ID numbers are re-mapped or removed to de-identify students within the ERC data warehouse. • Further requirements for preventing identification of individuals are set out in written confidentiality agreements signed by researchers—including a general guideline that any data cell with a composite size of less than five must be suppressed in any data released from the ERC. • Data users are required to go through training on protecting confidentiality and privacy.
5. How much time and cost are involved in requesting and receiving data?		• According to ERC website, under normal circumstances, review of requests will be completed in three to seven working days, with an expedited review possible upon special request. • Access to data through ERC client workstations must be scheduled on a first come, first served basis. • No cost for on-site access by researchers is set out.

The University of Texas at Austin Education Research Center

Confidentiality Agreement

between

The University of Texas at Austin Education Research Center and

_____

Researcher Name

for

_____

Research Project

As an associate of the Texas ERC, you have access to confidential data from the Texas Education Agency and Texas Higher Education Coordinating Board. By your initials and signature below, you acknowledge and agree:

_____1) that you have received a copy of both the Primary Contract (Interagency Cooperation Contract between The University of Texas at Austin, the Texas Education Agency, and the Texas Higher Education Coordinating Board) and The University of Texas at Austin Education Research Center Terms and Conditions for Using the Texas ERC,

_____2) to abide by the terms of the Primary Contract and the Terms and Conditions for Using the Texas ERC and its subordinate processes and procedures,

_____3) to access and use the Texas ERC data for only authorized research,

_____4) not to attempt to identify individuals or publicly release confidential data,

_____5) to ensure that all research conducted and all generated research products (papers, abstracts, PowerPoint presentations, etc.) using Texas ERC data are compliant with the Family Educational Rights and Privacy Act (FERPA),

_____6) never to remove unapproved confidential information from the physical or electronic workspace of the Texas ERC,

_____7) to request Texas ERC review and approval of all research products generated using confidential Texas ERC data, prior to any public release of those products,

_____8) to report, as soon as possible, any known or suspected breach of confidentiality, including the removal or inappropriate sharing of data, to the Director or Associate Director of the Texas ERC,

_____9) that access to the Texas ERC can be suspended based on any violation of this contract or risk of unauthorized disclosure of confidential information,

_____10) and grant permission for the manual and electronic collection and retention of security-related information, including photographic or videotape images, of your attempts to access the facility.

Signature: _____ Date: _____

Researcher

Signature: _____ Date: _____

Principal Investigator

Signature: _____ Date: _____

Dr. Pedro Reyes, Texas ERC Director

Signature: _____ Date: _____

Dr. Celeste Alexander, Texas ERC Associate Director

Tools for Monitoring the Health Care Safety Net
Integrated State Data Systems (continued)

Components for Successful Data Integration

Impetus for South Carolina's Integrated Data System

The need to answer complex policy questions was actually the catalyst for the integrated data system in South Carolina. In 1992, the following systemic and population-based policy issues facilitated data sharing activities:

- Improving access to health care services.
- Containing health care costs.
- Maintaining or improving existing quality of care in a cost-effective manner.
- Enhancing informed decisions in the selection of health care providers, facilities, and services.
- Determining the appropriate types of health care services needed for the State's growing elderly and disabled populations.
- Determining the effect of lifestyle, social, environmental, and genetic factors on health.
- Evaluating and improving the types of treatment being provided in a wide range of settings.

The data plan that was developed took each policy issue and identified the types of data needed for analysis. This information was then cross-referenced with data and systems that had been previously inventoried. This aided in determining which datasets were readily available and what changes would have to be made to existing systems to answer these policy questions.

Given the complexity of the policy issues, it became clear that the integration of data would be the only meaningful way to proceed. Integration of data would permit tracking cohorts from providers' offices through emergency departments and inpatient hospitalizations, as well as tracking clients across a range of State agency-based programs.

During the planning process, it was determined that the collection of personal identifiers was necessary to track populations across time and service providers. Without these identifiers, there was no way to link multiple datasets from multiple sources. Using identifiers to create a key linker or unique identification number allowed South Carolina to reach beyond "silo-oriented" data systems to achieve an enterprise integrated data system that contains rich information for all areas of health care and the public sector. An example of this is a study

conducted that examined health service utilization 6 months suicide was committed. It was determined that more than 30 percent of people who successfully committed suicide sought care from an emergency department or community mental health center 6 months before taking their own lives. This level of analysis would not have been possible without using the identifiers to create a key linker to integrate multiple datasets.

Technical Issues

Because of the ever-evolving nature of hardware and software, it is imperative that an integrated system not be built based on technical availability. Rather, it should reflect a research philosophy that transcends technology. There are several rules that South Carolina has adopted in developing the technical components of its integrated data system.

Technical Foundation for Integrating Data Is a Successful Key Linker System

South Carolina began "unduplicating" all personal identifiers on 1996 records. Each unduplicated person is randomly assigned a number generated by a computer algorithm. The number is not affiliated with any identifier associated with the individual, e.g., Social Security No. or date of birth. The number remains with that individual for all subsequent service use, regardless of data source or service provider. The algorithm accounts for misspellings, aliases, and name changes. For each additional record, public or private, submitted to the Office of Research and Statistics, a comparison is made to the "unduplicated" person file. If that individual is found, then the designated key linker is assigned to that episode of service. If that individual is not found, then he or she is added to the unduplicated file and assigned the next available number. To protect confidentiality, an individual's personal identifier is never associated with the service received. The final statistical file has no personal identifiers, only the key linker.

Develop Expertise in Dealing With Agency Datasets by Establishing Key Agency Contacts

No one knows the data better than personnel who work with the original data sources. It is important that the integrated data staff maintain an ongoing relationship with the key agency data staff. Administrative data systems evolve to meet the changing needs of the programs and services they administer, and the learning curve for the integrated data staff member never ends. A strong working relationship between the integrated data staff member and staff from the data source is essential.

Be Prepared to Deal With Longitudinal and Sometimes Duplicative Data

The planning process should include anticipating how to deal with longitudinal and duplicative data. Only by developing protocols for dealing with these issues will it be possible to present an accurate picture of the population studied.

Be Prepared to Address Transiency Issues

It has been South Carolina's experience that many of its lower-income residents do not reside in the same place for very long periods of time. A sophisticated GIS system can aid in tracking the movement of various populations. In structuring an integrated data system, it is important to build in an address-matching software component that compiles an address history file that allows tracking geographic mobility over time. This is especially helpful in studying

environmental linkages to certain health conditions such as lead poisoning and asthma.

Understand the Complexities Surrounding Agency Data Systems

It is not enough to understand the data that agencies and organizations share through an integrated data system. It is equally important to understand the data policies, computer hardware, and data utilization of each agency or organization. This knowledge will enrich the integrated data system's ability to analyze data.

Policies and Data Systems Change Within Agencies

Have a plan ready to address change in data policy and data systems. As data elements, formats, and variables change or are deleted, the successful integrated data system will experience little disruption because a plan has been implemented to absorb such changes.

Organizational Structure

The importance of a sound organizational structure cannot be overemphasized. Locating the integrated data system in a neutral organization diffuses power issues and allows data to be analyzed in a non-competitive and apolitical environment. In South Carolina, the Office of Research and Statistics (ORS) is responsible for integrating data and is located in the Budget and Control Board of the State government. The Budget and Control Board, which reports to both the Governor and Legislature, is a neutral body that does not provide services, nor does it oversee any services or other agencies. By being a neutral agency, ORS provides agencies with data that empower them to better serve the State's residents.

Building Trust

Depending on the environment, one of the most challenging components of an integrated data system is establishing trust among the various individuals and organizations contributing data. Data can personify power to many; thus, sharing data in a hostile environment can at best create insecurities, if not resistance, among data partners. South Carolina has spent decades building trusting relationships with data partners, and as a result has data from virtually every State agency, the private health care sector, many not-for-profit organizations, and medical clinics. South Carolina follows a set of general principles for building trust:

Do Not Try to Replace Existing Agency Functions

It is best to augment what is already occurring so that no one feels threatened. Most agencies and organizations have in place statistical and research teams that support their respective data efforts. When integrating data, it is best to assure the data partners that their existing functions will not be interrupted, and that the new integrated data system will not replace the work they are currently doing. It is best to identify ways in which the integrated data system can aid in or augment the existing work. The fear of being replaced is sufficient to sabotage the integrated data effort.

Each Organization (Public and Private) Is Treated Equally

Rules for partnering in an integrated data system must be consistent among all partners. To

provide any partner with preferential treatment erodes the very foundation upon which the system is built. Activities such as data access or exchange should be consistently applied across all the partners. To do otherwise creates a competitive environment that also may sabotage integrated data efforts.

Steer Clear of Egos and Political Issues

It is never wise to use the integrated data to harm partners. This is a philosophy that must be agreed upon by all partners in the initial phases of development. Using the integrated data as a weapon to do harm creates an atmosphere of threat, which results in the diminished capacity of the system.

Never Criticize an Agency/Organizational Data System

Criticism of a partner's system generates feelings of insecurity and inferiority, resulting in a partner's resistance to sharing data. The omission of a partner's data can have substantial effects on the overall value of the integrated data system.

Make the Data Extraction Process as Non-intrusive as Possible

Complementing rather than competing with a partner's existing system for data extraction creates a natural tendency for that partner to share data through the integrated data system. Participation in the system should be effortless for the partners. Creating additional work for participation generates resistance.

Do Not Duplicate or Compete With Agency Statistical Offices

An integrated data system should empower the agency with robust data to further its mission and reach its constituents. Coming between an agency and its funding source, governance, or clients will have a detrimental effect on the integrated data system. Taking attention away from the partners will sabotage the system.

Gain Support of Statistical and Information System Offices First

These individuals will be working intimately with the integrated system. Have their complete buy-in and support for the effort before approaching those at the executive level. These staff will be the best people to advocate to the executive level for participation in the integrated data system.

Secure Support for Data Sharing and Linkage First

Do not attempt to work out all the issues at the beginning. Given the many regulations surrounding privacy and confidentiality, it would be easy to lose focus battling over the finer points of the system. These types of discussions can bring an integrated data system to a halt before the first data file is loaded. It is imperative to reach a mutual agreement on the philosophy of the integrated data system first and revisit it often to reinforce a cooperative spirit. The details will work themselves out as long as everyone is in agreement on the goal.

Put It in Writing

Execute memoranda of understanding/agreement establishing rules of data sharing and

assuring agency control. Nothing demonstrates trust more than a written commitment. In an integrated data system, the partners should retain control over their own data. They should provide direction on what types of linkages should occur. Putting this commitment in writing assures trust. In South Carolina, partners have learned that the more they share, the more they get back in return. Integrated data systems are an exponential investment in knowledge.

Rules for Privacy, Confidentiality, and Data Access

Legislation is a mechanism for securing data sharing through an integrated data system, if it is pursued as a united effort by the partners. However, without the support of the partners, the system will not be successful, even if partners are mandated to share data. In South Carolina, original legislation addressed only private health care sector data. The legislation established the South Carolina Data Oversight Council, which is comprised of representatives from the private sector (hospitals, physicians, and nursing homes), public sector (Governor's office, Health Department, Health and Human Services Department), third-party payers, and the business community. The release of the private health care sector data is provided at the Council's direction, not that of the ORS. This reinforces one of the rules for building trust by treating everyone equally. This impartial body reviews data requests and releases data based on criteria as established in their policies and procedures.

Historically in South Carolina, release of public or State agency data is dictated by the agency providing the data through a Memorandum of Understanding or an internal data review committee or both, depending on the policies of the respective agencies. If a data request requires links of multiple datasets, ORS facilitates the requestor's proposal by bringing the partners together to review the data request. Previous experience has demonstrated that this fosters a collaborative spirit and maintains the partners' communication about and awareness of current research. It invariably results in a positive experience for the requestors and data partners.

In 2002, a proviso was passed to solidify data sharing activities among State agencies in South Carolina. The State law pre-empts HIPAA for collection of data. Key requirements include:

- Agencies collect and provide client data to the Budget & Control Board's ORS as a neutral organizational structure for integrating data.
- ORS establishes Memoranda of Understanding with each agency that specify confidentiality, release of data, etc.

Agencies retain ownership of their data, and no data are released by ORS without the express permission of the agency. The key linker system, which de-identifies individuals, fosters adherence to all Federal and State laws and regulations pertaining to confidentiality and privacy.

In addition to the laws and regulations specific to the integrated data system in South Carolina, there also is a statutory environment that facilitates the receipt of data from agencies. South Carolina laws permit the sharing of data, with identifiers, to registries such as those for cancer and reportable diseases. Sharing integrated data for this purpose ensures a more accurate registry, thus strengthening the services provided to individuals with these conditions.

Staff

Investment in statistical and research staff is essential, as they provide the analytical ability and thought processes that empower an integrated data system. Knowledge of data systems and

statistical prowess are benchmark skills necessary to develop and maintain such a system. Sufficient information technology skills are important for support roles, but the integrated data system in South Carolina was built primarily through the efforts of statisticians. The South Carolina team possesses a diverse background that includes former clinicians, human services managers, and teachers, as well as mathematicians and statisticians. The professional experiences of the team collectively enrich the analyses required for South Carolina's complex health and human services problems. South Carolina has trained its own generation of integrated data professionals. The real investment of integrated data systems is found in the people who develop and nurture them.

Conclusion

Existing administrative systems are data rich, and the workforce, given its technical skill and prowess, is capable of using these systems to generate significant change. The next frontier is deploying a community of managers, researchers, advocates and consumers with a research agenda that:

- Recognizes that health is not organizationally based nor is it captured in a "silo" data system.
- Focuses on understanding the tangential as well as the core determinants of wellness.

The tools—including the data, computer systems, software, and integration programs—are ready. Full data integration across the health and human services spectrum needs to be a top priority for policymakers and service providers to improve health care for vulnerable populations.

Current as of September 2003

Internet Citation:

Bailey WP. *Integrated State Data Systems*. Tools for Monitoring the Health Care Safety Net. September 2003. Agency for Healthcare Research and Quality, Rockville, MD.
http://www.ahrq.gov/data/safetynet/bailey.htm

DUA Toolkit

A guide to Data Use Agreements in the HMO Research Network

Purpose and Description

This guide was created to facilitate the establishment of Data Use Agreements (DUAs) for multi-site studies within the HMO Research Network. It includes information about:

- when DUAs are needed
- the steps involved in putting a DUA in place
- tools and resources related to DUAs and PHI disclosures
- best practices and common pitfalls

Comments or Questions

If you have questions about DUAs not sufficiently addressed here, refer to your local DUA contact person. Please refer to the table of site contacts and signatories.

If you have specific comments or feedback about this guide, please contact Ella Thompson at Thompson.e@ghc.org.

Acknowledgments

Group Health Center for Health Studies, HealthPartners Research Foundation, and Kaiser Permanente Northern California Division of Research led the development of this Toolkit on behalf of the HMO Research Network. This work was funded by the National Institutes of Health, under Contract No. HHSN268200425212C, "Re-Engineering the Clinical Research Enterprise."

TABLE OF CONTENTS

APPENDICES

WHAT IS A DATA USE AGREEMENT?

A Data Use Agreement (DUA) is an agreement that governs the sharing of data between research collaborators who are covered entities under the HIPAA privacy rule. A DUA establishes the ways in which the information in a limited data set may be used by the intended recipient, and how it is protected.

Advantages of a DUA

The HIPAA privacy rule allows a covered entity to use and disclose a limited data set (LDS) for research without obtaining an authorization or a waiver of authorization. A covered entity (e.g., a health plan) may disclose a LDS to another entity or researcher who is not a covered entity when a DUA is in place.

Important up front considerations

1) Expect that analyses and manuscript authorship will be spread across sites, and ensure all potential authors will have access to data.

DUAs ARE ALWAYS STUDY SPECIFIC.

Blanket DUAs do not exist between organizations.

Permissions outlined in a DUA

1) Who may receive and use the limited data set
2) Allowable uses and disclosures by the recipient

Assurances outlined in a DUA

1) The recipient will not try to identify or contact subjects represented in the LDS.
2) The recipient will not use or disclose/share the data in ways other than stated in the agreement, or as otherwise required by law.
3) The recipient will safeguard the data to prevent such misuse or unauthorized disclosures.
4) The recipient will report any misuse or unauthorized disclosure as soon as known.
5) The recipient will ensure that any agents, including subcontractors, agree and are bound to the restrictions and conditions of the DUA.

WHEN DO I NEED A DUA?

To put it simply, you need a DUA anytime you are sharing data that are not de-identified in a manner that was not explicitly covered in the consent form. Sharing a de-identified data set does not require a DUA, but limited data sets may be shared only after a DUA is in place. The first step is to determine what type of data set you are working with.

Do I have a limited data set?

Limited data sets do NOT include direct identifiers (like name and address). However, limited data sets MAY contain the following indirect identifiers:

- town or city, state, zip code;
- ages in years up to 90 years (must aggregate all ages 90 or older);
- dates directly related to an individual – such as birth date, date of death, admission date, discharge date, visit date, diagnosis date, etc. *(Month/year is preferred).*

A unique study ID can be included in both limited and de-identified data sets – but the number can NOT be an encoded identifier, such as a scrambled birth date, patient initials, last four of social security number, and so on.

Working out the terms of a DUA sometimes takes more time and effort than foreseen.

CONSIDERATION

Is aggregated data or a de-identified data set an option for your study?

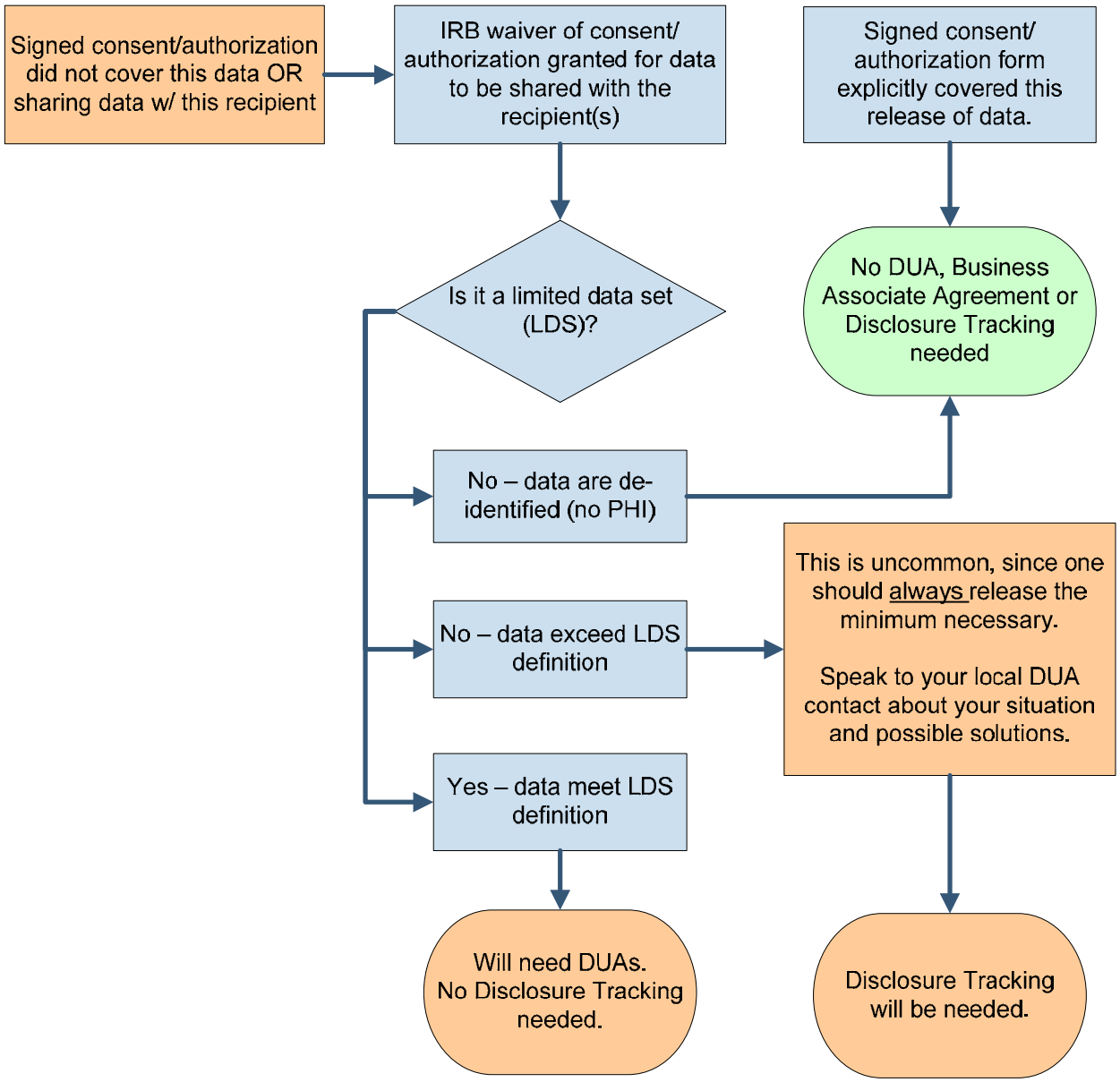
Do I have a de-identified data set?

Data are considered de-identified if there's no reasonable way they could be used to identify a person. Thus, de-identified data sets may NOT contain any of the following protected health information (PHI):

- name
- social security number
- geographic subdivisions smaller than state (e.g., street address, city, county, precinct, zip code, and their equivalent geo-codes)
- all month and day elements of dates directly related to an individual (e.g., birth date, admission date, discharge date, date of death, etc.)
- all ages AND/OR all birth month/day/year elements for persons over 89 years (data may be aggregated into a single category, 'age 90 or older')
- voice and fax telephone numbers
- electronic mail addresses
- medical record numbers, health plan beneficiary numbers, or other health plan account numbers
- certificate/license numbers

- vehicle identifiers and serial numbers, including license plate numbers
- device identifiers and serial numbers
- Internet Protocol (IP) address numbers and Universal Resource Locators (URLs)
- biometric identifiers, including finger and voice prints
- full face photographic images and any comparable images
- any other unique identifying number, characteristic, or code (Note: voice recordings and lab specimen accession numbers are considered PHI.)

Flow Diagram: Do I need a DUA?



My data set exceeds a limited data set--What now?

Remember to release only the Minimum Necessary. If you do NOT have a signed written consent authorizing data sharing with the recipient AND you exceed the definition of LDS:

1) Obtain an IRB Waiver of Authorization.

2) Work out contractual solutions with your site administrators (e.g., Business Associate Agreement (BAA), Memorandum of Understanding (MOU), Non-disclosure Agreement, etc.).

3) Specify both the patients and type of PHI sent outside your institution according to your health plan's Disclosure Tracking procedures.

Step 1 and/or 2 may require a great deal of time and resources.

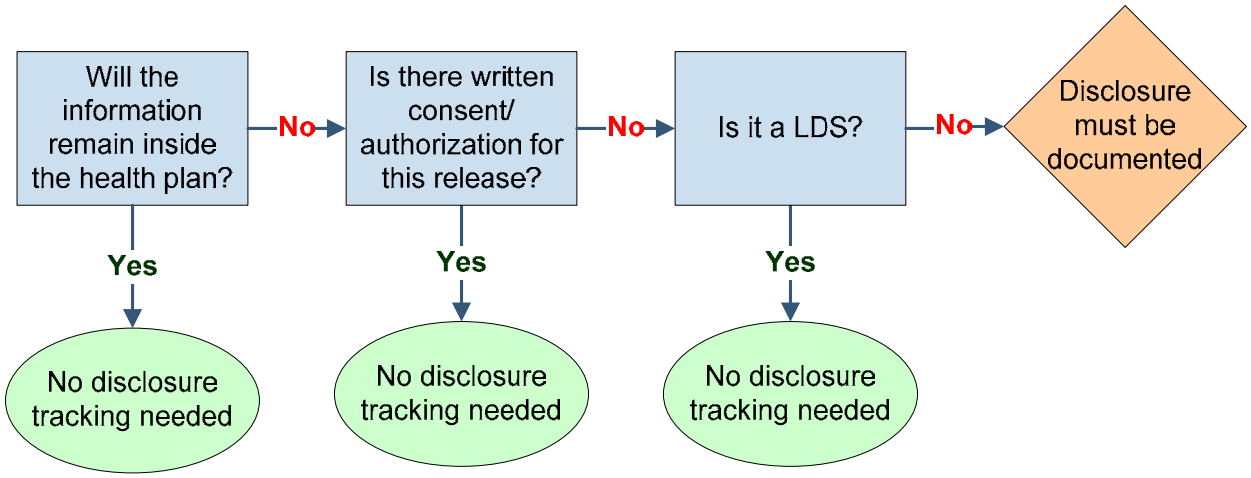
CONSIDERATION

Is it possible to alter your analysis plan so only a LDS is sent?

Disclosure tracking

Disclosures must be tracked any time protected health information is disclosed and either of the following apply:

- Authorization or a waiver of authorization has not been granted.
- Data exceed the definition of a limited data set.



SETTING UP A DUA

There are three important steps to follow when setting up a DUA:

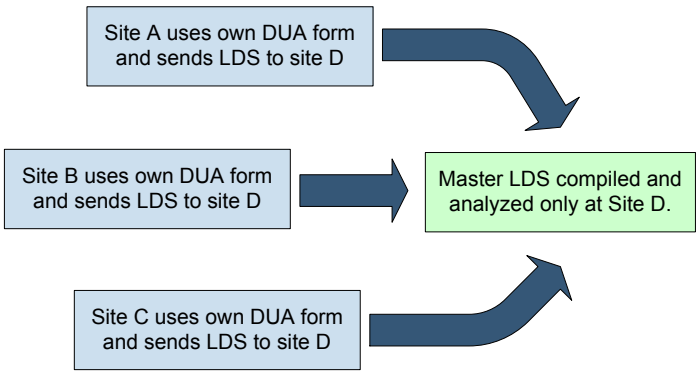
1) Identify the DUAs that are needed.
2) Build from a template or previous DUA.
3) Finalize the paperwork.

Step 1: Identify the DUAs that are needed

To help illustrate this step, consider the following three common scenarios:

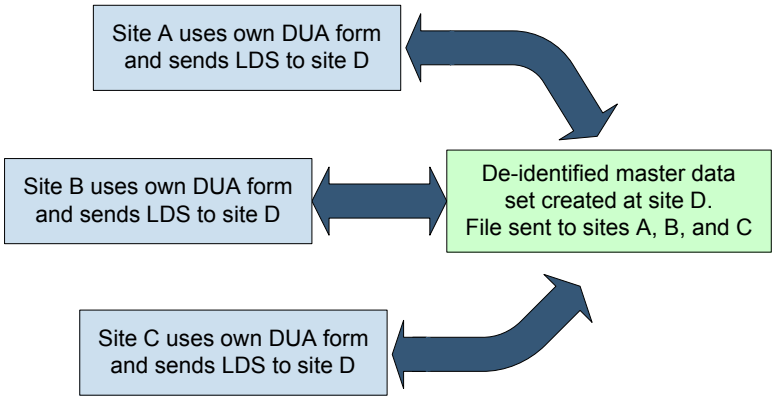
Scenario 1

Sites A, B, C, and D have all collected data in a multi-site study. Site D will create and analyze a master limited data set (LDS), but will NOT send the LDS back to the other sites. Sites A, B and C need to establish a DUA with site D. Each site will use its own form or an agreed upon template DUA.



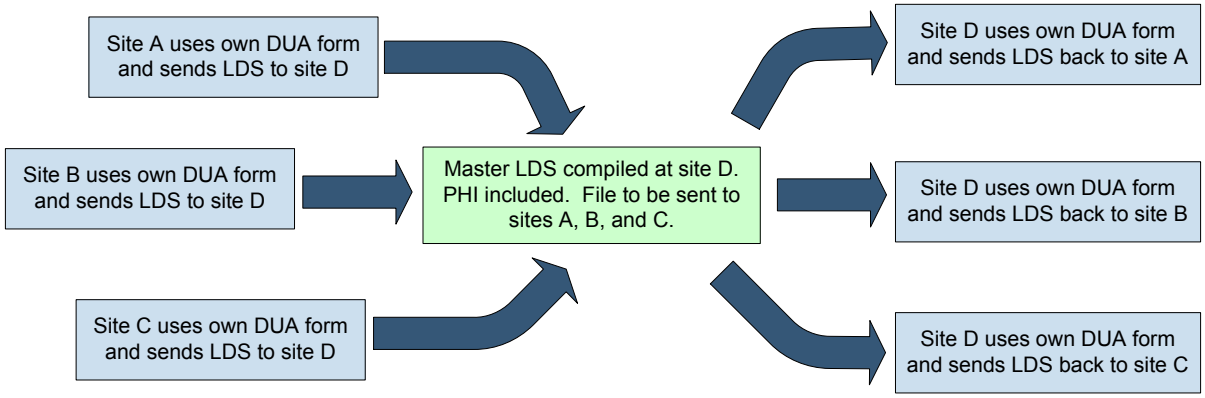
Scenario 2

This scenario is the same as Scenario 1 above, except that site D will compile the LDS and then create a de-identified data set (no PHI) to send back to the other sites for local analyses. Because the data being sent back to sites A, B, and C has been de-identified, no new DUAs are needed.



Scenario 3

This scenario is the same as Scenario 2 above, except that site D will compile the LDS with PHI included to send back to the other sites for local analyses. Site D needs DUAs with A, B and C before the new LDS can be sent. Site D will use its own DUA form since they house the new master LDS being sent.



Step 2: Build from a template or previous DUA

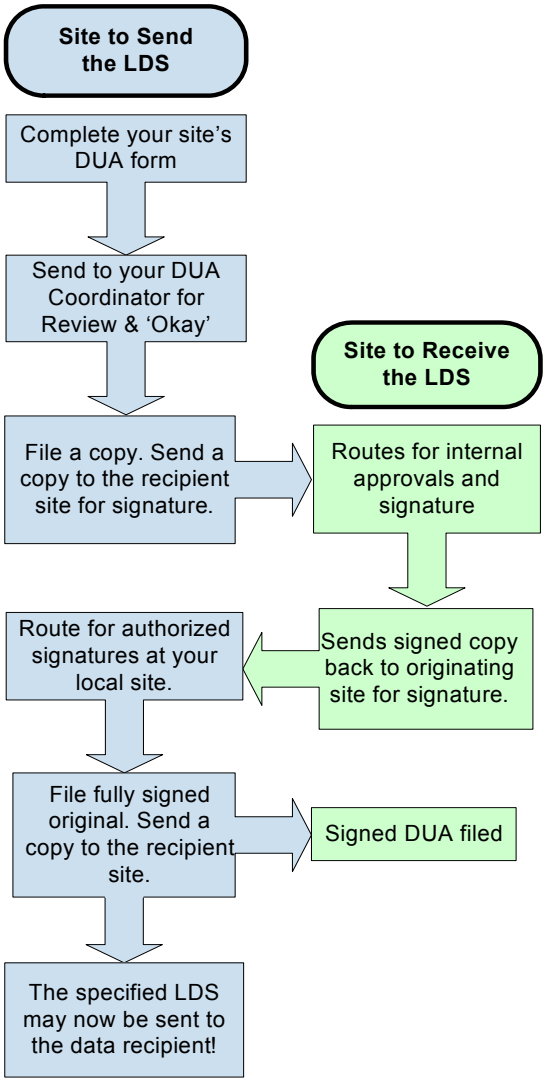
Most HMORN institutions already have a DUA template. Check with your site DUA coordinator to find out if your research center has a template. Templates differ a bit based on the individual institution's legal department.

It can also be helpful to find past or current DUAs between your institution and the recipient. These may provide useful precedents.

Refer to the table of DUA signatories, point persons and contacts.

Step 3: Finalize the paperwork

The diagram at right shows the typical flow of paperwork within and between sites setting up a DUA.



PROACTIVELY PLANNING FOR SUCCESS

Tips and best practices

1) Learn the process for setting up DUAs at your own institution including who has authority to sign.

2) Find out if your site had a previous DUA with the proposed recipient. You may be able to use that agreement as a template or for precedent language.

3) Ensure as much time as possible to allow for interpretation and possible reaction to legal wording in the agreements. Set your DUAs up early in the life of the project.

4) Ensure all authors will have access to data. Anticipate opportunities to spread analyses and manuscript authorship broadly across sites and write the DUAs to reflect this.

5) Follow communication pathways set up at individual sites. Circumventing the process causes confusion and adds time.

6) Sync up language in the contract with DUA-related terms. If issues are already addressed in the subcontract, time and resources can be saved downstream.

7) Clarify specific data elements needed for the analysis up front.

8) Required components of a DUA are spelled out in HIPAA. Avoid using a DUA to insert additional requirements more appropriate for a contract.

9) Keep the following documents in the project files at each site:

    - Fully signed DUA.
    - Documentation of content of the data sent/received (e.g., SAS proc contents report).
    - Cover letter or email documenting data transfer.

Issues commonly leading to delays

Variations in expectations and practices at the local level are a factor in every multi-site study. It can help both Investigators and Project Managers to be aware of the types of problems encountered by others.

- The DUA was written narrowly and uni-directionally. It did not account for the possibility of new analytic plans. For example, only the prime site could send pooled data to subcontractors. The DUA did not address sub-to-sub data sharing for secondary analyses, etc, or the addition of a new site.

- Local interpretation of regulations by legal counsel, etc. varied across sites, making mutual agreement much more difficult.
    - Agreement on which state (or site) has jurisdiction, should disputes arise.
    - One site may require more stringent security protections than another site.

- State laws prohibited sites from reaching mutual agreement on some DUA terms.
    - Minnesota, Washington, and Oregon all have state laws pertaining to certain types of data (e.g., the Oregon Genetic Privacy Law) which may supersede federal regulations in the HIPAA Privacy Rule.

- Receipt of aggregated summary data only may preclude certain analyses.

- Sites may hesitate to stray from language used in past DUAs or may not want to make changes to a pre-approved template.

- Trying to involve a non-HMORN based Investigator or business associates prolonged negotiations.
    - Example: Data collection or data entry service

- Sites have differing views on the degree of assumed risk to the health plan (e.g, in the event of an unauthorized disclosure) when data are shared.
    - Example: Some health plans may view quality of care data as being a greater risk that data on use of preventive services.

APPENDICES

More about PHI and Data Disclosure

Protected health information is defined under HIPAA as *individually identifiable* health information. *Identifiable* refers to data explicitly linked to a particular individual as well with data that could enable individual identification. Identifiers include obvious ones like name and Social Security number. Others are:

- all geographic subdivisions smaller than a state, including street address, city, county, precinct, Zip Code, and their equivalent geo-codes
- all elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
- voice and fax telephone numbers
- electronic mail addresses
- medical record numbers, health plan beneficiary numbers, or other health plan account numbers
- certificate/license numbers
- vehicle identifiers and serial numbers, including license plate numbers
- device identifiers and serial numbers
- Internet Protocol (IP) address numbers and Universal Resource Locators (URLs)
- biometric identifiers, including finger and voice prints
- full face photographic images and any comparable images
- any other unique identifying number, characteristic, or code

Under HIPAA's “safe harbor” standard, information is considered de-identified if all of the above have been removed, *and* there is no reasonable basis to believe that the remaining information could be used to identify a person.
As an alternative to using fully de-identified information, HIPAA makes provisions for a limited data set for which direct identifiers (like name and address) have been removed. but not indirect ones (such as age).

A *limited data set* may not include any individually identifiable information (PHI) except for the following elements, subject to the minimum necessary standard:

- Town or city, state, and zip code

- Any date directly related to an individual (such as: birth date*, admission date, discharge date, date of death, visit date, diagnosis date….

- Any ages over 89

Both limited data sets and de-identified data sets may include a study number assigned for the project as long as it is not a combination of numbers that would allow identification of the individual (such as a scrambled birth date and Social Security Number).

Under HIPAA, the general rule is that researchers must have valid *authorization for all uses and disclosures of PHI* in connection with research.

- "Protected health information (PHI)" means individually identifiable health information transmitted or maintained in any form or medium.

- "Use" means, with respect to individually identifiable health information, the sharing, employment, application, utilization, examination or analysis of such information *within the entity that maintains such information*.

- "Disclosure" means the releases, transfer, provision of access to, or divulging in any manner of information outside the entity holding the information. A valid research authorization must be in writing, must be signed by an individual, and must contain certain elements.

A valid authorization must include specific elements:

- A description of the PHI being used

- A statement of the purpose of the use of PHI

- A list of those who can use the PHI

- A list of those who can receive the PHI, including the possibility of re-disclosure

- Information about the expiration of the authorization

- Information about the right to revoke the authorization

If an actual expiration date is not provided, then a note pointing this out is required. A statement explaining an expiration event such as the end of the research project is also acceptable.

As to the right to revoke, the authorization must either explain that right or refer to the covered entity's privacy notice, if that is applicable. A revocation must be in writing and can be made at any time, but it may not be effective if a research study has already relied on the authorization. This reliance element only affects information gathered before the revocation and does not allow the entity to disclose PHI after the revocation occurs.

The covered entity – that is, “health plans, health providers and health clearinghouses” or “any entity in the health sector that uses health information in the regular course of business” – may require the authorization as a condition of providing research-related treatment. In general, authorizations may not be combined with other documents, such as the notice of privacy practices or an optional consent, that is, a document signed

If a *limited data set* will be released outside of your health plan or accessed/used by anyone not employed by your health plan without a signed authorization or consent form of each individual whose data are used, then documentation of an IRB waiver of authorization must be kept on file by project staff and a DUA signed by the recipient of the data may be required.

If any *PHI beyond a limited data set* will be released outside of your health plan or accessed/used by anyone not employed by your health plan without a written authorization signed by each subject whose data are used, then documentation of an IRB waiver of authorization must kept by project staff and project staff must enter pertinent data into a disclosure tracking file at your health plan. In addition, a business associate agreement may be required.

Links to Additional Resources

National Institute of Health (NIH) – HIPAA Information for Researchers
Educational materials listed on the NIH site are readable and complete. Several address limited data sets and DUA issues and include FAQ at the end:

Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule

Clinical research and the HIPAA Privacy Rule

Health services research and the HIPAA Privacy Rule

Research Repositories, Databases and the HIPAA Privacy Rule

HIPAAlert Newsletter

HIPAAlert is a monthly independent newsletter that provides coverage of major HIPAA-related developments. The newsletter features expert commentary, case studies, Q&A, compliance tips, links to original, full-text documents and helpful HIPAA resources.

Archived newsletters Subscriptions

Duke IRB website

The Institutional Review Board of Duke University has a very thorough website covering many aspects of research compliance including HIPAA.

HIPAA handbooks

The HIPAA Training Handbook for Researchers, by Lawrence Muhlbaier

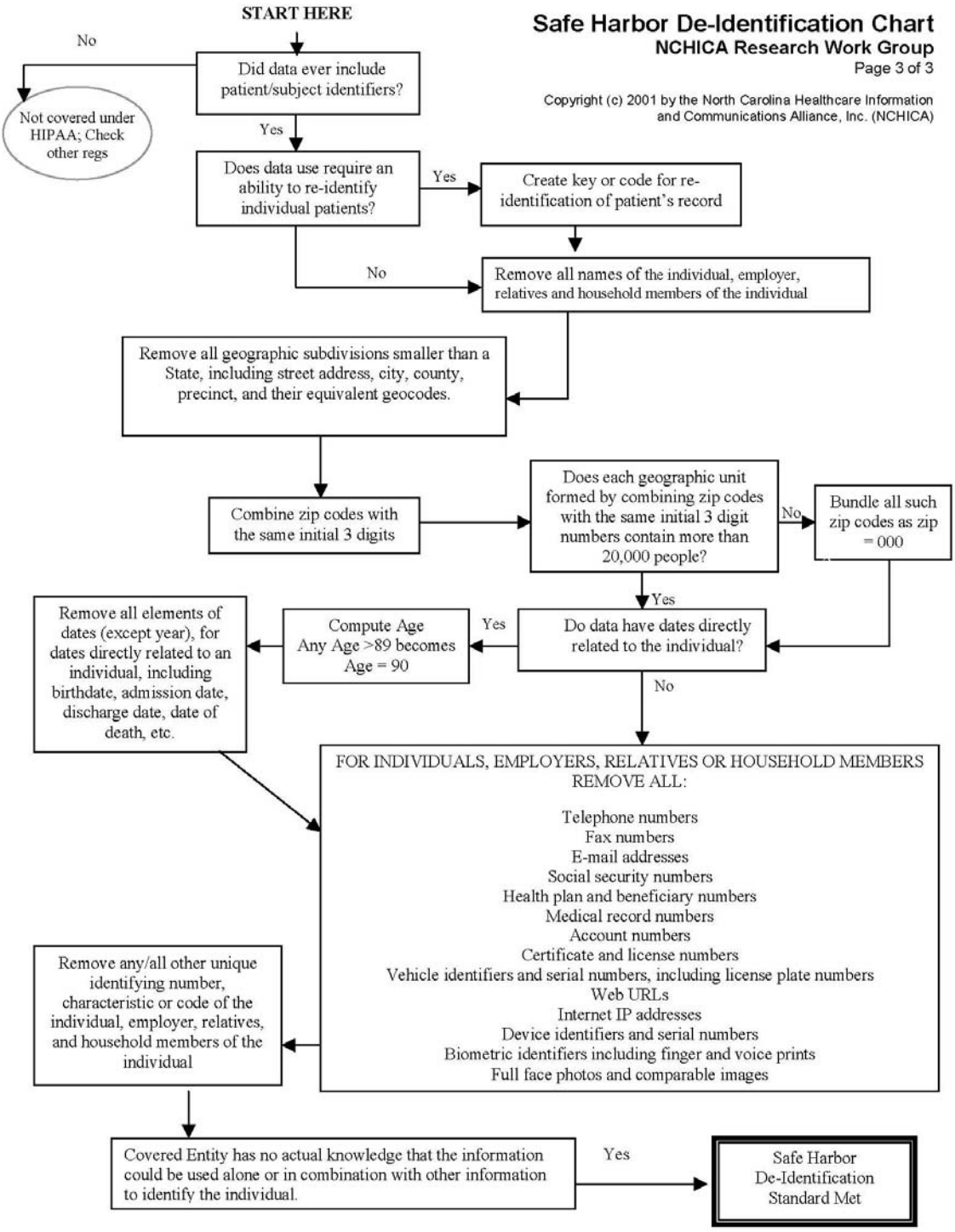
Dr. Muhlbaier is a statistician at Duke University and a member of the Duke IRB. This handbook provides an overview of HIPAA as it pertains to research. Handbooks may be ordered in sets of 15 books at $4 each at www.hcmarketplace.com.

HIPAA in Clinical Trials: A Practical Guide for Research Compliance, by Lawrence Muhlbaier.

At 151 pages, this is a more thorough discussion of HIPAA rules as they pertain to research. There are valuable tips and useful interpretation. State laws and interpretations by your local health plan and other research partners are sometimes at odds with Dr. Muhlbaier's suggestions, so this should not be viewed as the last word. But it is thoughtfully written and helpful for gaining understanding of the complexities involved. Available from www.hcmarketplace.com for $199 each.

Safe Harbor De-Identification Chart and Other NCHICA Tools

The North Carolina Healthcare Information and Communications Alliance's (NCHICA) HIPAA Implementation Planning Task Force produced the de-identification chart below and many other useful documents and tools to educate the healthcare community about HIPAA – available at http://www.nchica.org/HIPAAResources/Samples/Portal.asp.



Frequently Asked Questions

Covered Entities

A covered entity is a health care provider that conducts certain transactions in electronic form, a health care clearinghouse, or a health plan. A simple way to check if an institution is or is not a covered entity is to look for their HIPAA Notice of Privacy Practices (NOP) on the internet. Covered entities are required to display their NOP.

Is the Center for Disease Control a covered entity?

No. Although the CDC collects clearinghouse-like data, it is not an agency that handles treatment, payment, and referral transactions for health care providers.

Is the CSS/SEER (the Cancer Surveillance System) a covered entity?

CSS/SEER is a Public Health Authority –that is, an agency of the government that is responsible for public health matters as part of its official mandate. The FDA and OSHA are also Public Health Authorities. HIPAA permits covered entities to disclose protected health information, without authorization, to other PHAs. Click here for more info.

Are covered entities required to document incidental disclosures permitted by the HIPAA Privacy Rule, in an accounting of disclosures provided to an individual?

No. The Privacy Rule includes a specific exception from the accounting standard for incidental disclosures permitted by the Rule. See 45 CFR 164.528(a)(1).

Is it De-identified data?

May information de-identified under the Privacy Rule's "safe-harbor" method contain a data element that identifies a time period of less than a year (e.g., the fourth quarter of a specific year)?

No. The Privacy Rule's "safe-harbor" method for de-identifying health information requires removal of, among other elements, all elements of dates directly related to an individual, except for year. Thus, a data element such as the fourth quarter of a specified year must be removed if a covered entity intends to de-identify data using the "safe-harbor" method. See section 164.514(b)(1) of the Privacy Rule. From: NIH website

What lab data variables are permitted in a de-identified data set?

De-identified data cannot contain a lab accession number since they can be linked to consumer numbers in health plan data systems. Specimen collection and test dates are not permitted. Considered de-identified are: the year of the date, and the patient's age at the time of the test.

Can a de-identified data set contain an adverse event date?

No. De-identified data sets can contain only the year of the date, not the month or day. However, under HIPAA, there are special considerations for reporting adverse events. If your sponsor is the FDA, you may report adverse events without specific agreements. The minimum necessary standard applies. This would count as a disclosure and would need to be tracked.

Can I send aggregate data without identifiers or dates to a collaborator without putting a DUA in place?

Yes, provided that the likelihood of an individual being re-identified is small. For example if the number in each cell is significant, the data can be shared with other researchers. Even with very small number in a cell, the data may be safe to send, for example if the categories it represents are broad enough, e.g. ages in five- or ten-year groups.

Can I substitute the number of days between a date variable and another date (e.g. randomization date) for the full date of an event to de-identify or limit your data?

Yes, this is one way to de-identify data. But the recipient cannot have the reference date or other information enabling reconstruction of the actual dates. For example, permissible data to send for an immunization study might by Age-in-days-at-MMR#1, Age-in-days-at-MMR#2, and Age-in-days-at-RashDx. This would allow researchers to see if the RashDx occurred within a short time of the vaccinations without ever giving birth date or service dates. If handled in this way, data would be de-identified and could be sent without setting up a DUA. However, if the recipient already has data that would allow him to create dates from the information you send, then you are in fact sending a LDS (even if in piecemeal fashion) and so you would need to set up a DUA.

Is it Disclosure?

An external investigator would like to see paper questionnaires to do some data cleaning. There isn't any personal identifier information on the questionnaires, only a study number. Is there any reason not to send the questionnaires?

Your action will depend upon which variables are on the questionnaire and whether consents & authorizations are in place. Even though the data are on paper, it is still a dataset. A DUA could be required. Always check your IRB arrangements before releasing any data that are not de-identified.

If I share provider survey data, is it considered a disclosure?

No, as long as the data do not contain health information. Most provider surveys reflect the beliefs and practices of the provider and are therefore not health information. However, provider surveys may contain sensitive data, so check your IRB arrangements before releasing any data that are not de-identified.

Do we need to account for disclosures of updated contact information on study participants who gave oral consent before April 14, 2003?

It depends. In future studies such disclosures should be tracked. However, in an established study where regular contact with the participant has been maintained, this is not considered a disclosure. An important factor is whether the participants signed a HIPAA authorization form with your health plan - which are sometimes more stringent than HIPAA in categorizing interview results as PHI.

My study includes some subjects who are not health plan enrollees. We have disclosed some PHI on them. Do we need to account for such disclosures?

Yes, you are obligated to account for disclosures of PHI regardless of whether the data pertain to enrollees or other subjects. Because non-enrolled subjects do not have a consumer number in your health plan, there is not a way to capture individual level disclosures. Your site's disclosure tracking system should have a flag of some kind to mark such disclosures.

Glossary of Terms Used

Refer to the Privacy Rule on NIH's website for a complete listing of terms and their specific definitions.

Accounting for Disclosures - Information that describes a covered entity's disclosures of PHI other than for treatment, payment, and health care operations; disclosures made with Authorization; and certain other limited disclosures. For those categories of disclosures that need to be in the accounting, the accounting must include disclosures that have occurred during the 6 years (or a shorter time period at the request of the individual) prior to the date of the request for an accounting. However, PHI disclosures made before the compliance date for a covered entity are not part of the accounting requirement.

Authorization - An individual's written permission to allow a covered entity to use or disclose specified PHI for a particular purpose. Except as otherwise permitted by the Rule, a covered entity may not use or disclose PHI for research purposes without a valid Authorization.

Business Associate - A person or entity who, on behalf of a covered entity, performs or assists in performance of a function or activity involving the use or disclosure of individually identifiable health information, such as data analysis, claims processing or administration, utilization review, and quality assurance reviews, or any other function or activity regulated by the HIPAA Administrative Simplification Rules, including the Privacy Rule. Business associates are also persons or entities performing legal, actuarial, accounting, consulting, data aggregation, management, administrative, accreditation, or financial services to or for a covered entity where performing those services involves disclosure of individually identifiable health information by the covered entity or another business associate of the covered entity to that person or entity. A member of a covered entity's workforce is not one of its business associates. A covered entity may be a business associate of another covered entity.

Covered Entity - A health plan, a health care clearinghouse, or a health care provider who transmits health information in electronic form in connection with a transaction for which HHS has adopted a standard.

Data Use Agreement- An agreement into which the covered entity enters with the intended recipient of a limited data set that establishes the ways in which the information in the limited data set may be used and how it will be protected.

Disclosure - The release, transfer, access to, or divulging of information in any other manner outside the entity holding the information.

Health Care Clearinghouse - A public or private entity, including a billing service, repricing company, community health management information system or community health information system, and "value-added" networks and switches that either process or facilitate the processing of health information received from another entity in a nonstandard format or containing nonstandard data content into standard data elements or a standard transaction, or receive a standard transaction from another entity and process or facilitate the processing of health information into a nonstandard format or nonstandard data content for the receiving entity.

Health Care Provider - A provider of services (as defined in section 1861(u) of the Act, 42 U.S.C. 1395x(u)), a provider of medical or health services (as defined in section 1861(s) of the Act, 42 U.S.C. 1395x(s)), and any other person or organization who furnishes, bills, or is paid for health care in the normal course of business.

Health Information - Any information, whether oral or recorded in any form or medium, that (1) is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (2) relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual.

Health Insurance Portability and Accountability Act of 1996 (HIPAA) - This Act requires, among other things, under the Administrative Simplification subtitle, the adoption of standards, including standards for protecting the privacy of individually identifiable health information.

Health Plan - For the purposes of Title II of HIPAA, an individual or group plan that provides or pays the cost of medical care (as defined in section 2791(a)(2) of the PHS Act, 42 U.S.C. 300gg-91(a)(2)) and including entities and government programs listed in the Rule. Health plan excludes: (1) any policy, plan, or program to the extent that it provides, or pays for the cost of, excepted benefits that are listed in section 2791(c)(1) of the PHS Act, 42 U.S.C. 300gg-91(c)(1); and (2) a government-funded program (unless otherwise included at section 160.103 of HIPAA) whose principal purpose is other than providing, or paying for the cost of, health care or whose principal activity is the direct provision of health care to persons or the making of grants to fund the direct provision of health care to persons.

HHS Protection of Human Subjects Regulations - Regulations intended to protect the rights and welfare of human subjects involved in research conducted or supported by HHS. The HHS regulations include the Federal Policy for the Protection of Human Subjects, effective August 19, 1991, and provide additional protections for pregnant women, fetuses, neonates, prisoners, and children involved in research. The HHS regulations can be found at Title 45 of the **Code of Federal Regulations**, Part 46.

Hybrid Entity - A single legal entity that is a covered entity, performs business activities that include both covered and noncovered functions, and designates its health care components as provided in the Privacy Rule. If a covered entity is a hybrid entity, the Privacy Rule generally applies only to its designated health care components. However, non-health care components of a hybrid entity may be business associates of one or more of its health care components, depending on the nature of their relationship.

Individually Identifiable Health Information - Information that is a subset of health information, including demographic information collected from an individual, and (1) is created or received by a health care provider, health plan, employer, or health care clearinghouse; and (2) relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual; and (a) that identifies the individual; or (b) with respect to which there is a reasonable basis to believe the information can be used to identify the individual.

Institutional Review Board (IRB) - An IRB can be used to review and approve a researcher's request to waive or alter the Privacy Rule's requirements for an Authorization. The Privacy Rule does not alter the membership, functions and operations, and review and approval procedures of an IRB regarding the protection of human subjects established by other Federal requirements.

Limited Data Set - Refers to PHI that excludes 16 categories of direct identifiers and may be used or disclosed, for purposes of research, public health, or health care operations, without obtaining either an individual's Authorization or a waiver or an alteration of Authorization for its use and disclosure, with a data use agreement.

Minimum Necessary - The least information reasonably necessary to accomplish the intended purpose of the use, disclosure, or request. Unless an exception applies, this standard applies to a covered entity when using or disclosing PHI or when requesting PHI from another covered entity. A covered entity that is using or disclosing PHI for research without Authorization must make reasonable efforts to limit PHI to the minimum necessary. A covered entity may rely, if reasonable under the circumstances, on documentation of IRB or Privacy Board approval or other appropriate representations and documentation under section 164.512(i) as establishing that the request for protected health information for the research meets the minimum necessary requirements.

Protected Health Information - PHI is individually identifiable health information transmitted by electronic media, maintained in electronic media, or transmitted or maintained in any other form or medium. PHI excludes education records covered by the Family Educational Rights and Privacy Act, as amended, 20 U.S.C. 1232g, records described at 20 U.S.C. 1232g(a)(4)(B)(iv), and employment records held by a covered entity in its role as employer.

Use - With respect to individually identifiable health information, the sharing, employment, application, utilization, examination, or analysis of such information within the entity that maintains such information.

Waiver or Alteration of Authorization - The documentation that the covered entity obtains from a researcher or an IRB or a Privacy Board that states that the IRB or Privacy Board has waived or altered the Privacy Rule's requirement that an individual must authorize a covered entity to use or disclose the individual's PHI for research purposes.

Workforce - Employees, volunteers, trainees, and other persons whose conduct, in the performance of work for a covered entity, is under the direct control of the covered entity, whether or not they are paid by the covered entity.